Towards a Bulgarian Historical Newspaper Corpus – Construction of a Reading Order over the Text in Searchable PDFs

Nikolay Paev, Stefan Marinov, Ivan Kratchanov, Petya Osenova, Kiril Simov
IICT – BAS, Bulgaria
National Library "Ivan Vazov" – Plovdiv, Bulgaria

Introduction

- Old periodicals are an important source of information.
- Support research in many areas of social sciences and humanities (SS&H).
- The research in SS&H is in the stage of its active digitalization period, in which physical artifacts in archaeology, libraries, archives, museums, art galleries, and others are converted into digital objects.
- In parallel to digitalization of the artifacts, the methods for linguistic processing have also been developed – *opening new possibilities for research and cultural heritage preservation.*

Introduction

- We present the first steps in the digitalization and corpus creation of old Bulgarian newspapers in the period from the reinstatement of Bulgaria (1878) to the mid-20th century.
- Our focus is the handling of the **reading order** in the articles which is not a trivial task.

Reading order of old Bulgarian newspapers

- The included period is one of **great dynamics** in publishing, technology, and **typographic style** in Bulgaria.


Год. I

ПЛОВДИВЪ, Сѫбота 14 Юний 1908 год.

Брой 5.

6-ий СЕПТЕМВРИЙ

ОРГАНЪ НА СТАМБОЛОВИСТИТѢ.

Ив. Андоновъ
АДВОКАТЪ

ИЗТОЧНИ ЖЕЛѢЗНИЦИ.

ИЗВѢСТИЕ.

ИЗТОЧНИ ЖЕЛѢЗНИЦИ.

ИЗВѢСТИЕ.

ПЪРВА
БЪЛГАРСКА ФАБРИКА ЗА ФАБРИКУВАНИЕ НА МАРСИЛСКИ КЕРЕМИДИ
въ Т.-Пазарджикъ.

МАРСИЛСКИ КЕРЕМИДИ,

ЦѢНИ ИЗНОСНИ.

Адресъ: Лаки Галовъ
(Т.-Пазарджикъ).

МЛАДЕЖЪ
търси да прѣподава по
АНГЛИЙСКИ ЕЗИКЪ,
споразумение въ редакцията на
в. „Развитъ".

ВОДЕНИЧНИ КАМЪНИ
СОЛУНСКИ

А. Евстатиевъ
на Пловдивската гара.

МАКЕДОНСКИЯ ВЪПРОСЪ
И РАЗРѢШЕНИЕТО МУ.

1-ва БЪЛГАРСКА СПИРТНА ФАБРИКА И БИРАРИЯ
„СОЛУНЪ"

1-ва Българска Спиртна фабрика „Солунъ"

1-ва БЪЛГ. СПИРТНА ФАБРИКА И БИРАРИЯ „СОЛУНЪ"!

“6th of September” - 4 columns

Variety of reading orders


Година първа
Пловдивъ, 15 ноемврий 1911 година
Безплатенъ брой
Демократическо Слово

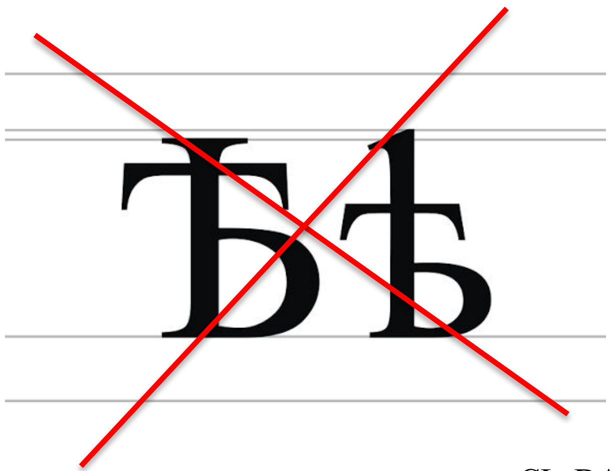
“Democratic speech” - 6 columns

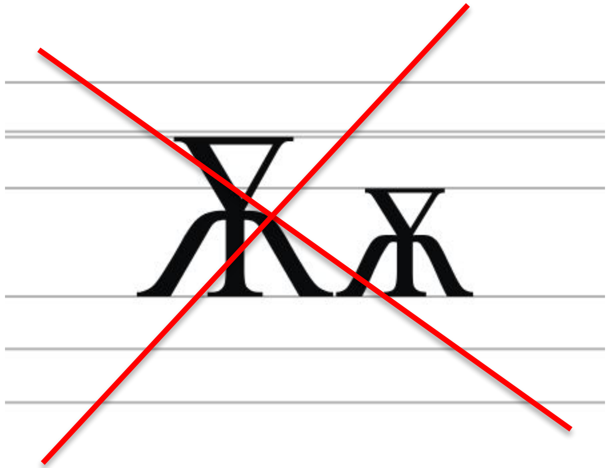

Стр. 2
„Бич“
Брой 1
Злободневка
ХРОНИКА
Стара приказка.

“Whip” - 5 columns

Bulgarian spelling system

- Spelling standards before 1945 have not been well-established. From 1870 to 1945, **eight spelling forms** were proposed and applied in various ways – officially, or non-officially.
- After 1945 Bulgarian **lost some letters**, among which the ‘yat’, the ‘ers’ in the end of the words, the nasal ‘big yus’.





Preparation of a Searchable PDF

- The newspapers chosen for the dataset, in their original paper form, are stored in the "Ivan Vazov" National Library - Plovdiv.
- Therefore, the OCR quality is strongly affected by the historical alphabets and spelling conventions, the archaic vocabulary and mixed-language content, as well as the limited dictionary coverage.
- A method for improving the accuracy of the OCR, provided by ABBYY FineReader, allows a targeted “training” for hard-to-recognize characters.

Preparation of a Searchable PDF

ПОЛЕЗНИ СЪВѢТИ.

Какъ да се държи мѣсо прѣсно. — Нарѣзва се месото на късове по единъ килограммъ, па се турятъ въ делви или каци, насипватъ на тѣхъ соль и ги покриватъ отгорѣ съ ситенъ прахъ отъ вжглища до два пръста дебело. По той начинъ може да се запази месото прѣсно до 15 дни, колкото и да е горѣщо врѣмето.

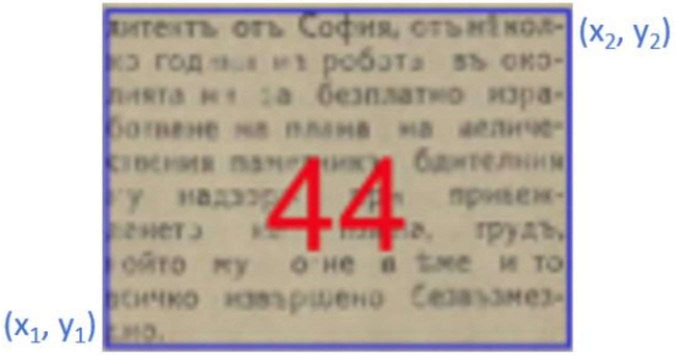
ПЪРВАТА НИ НАРОДНА ИЗЛОЖБА.

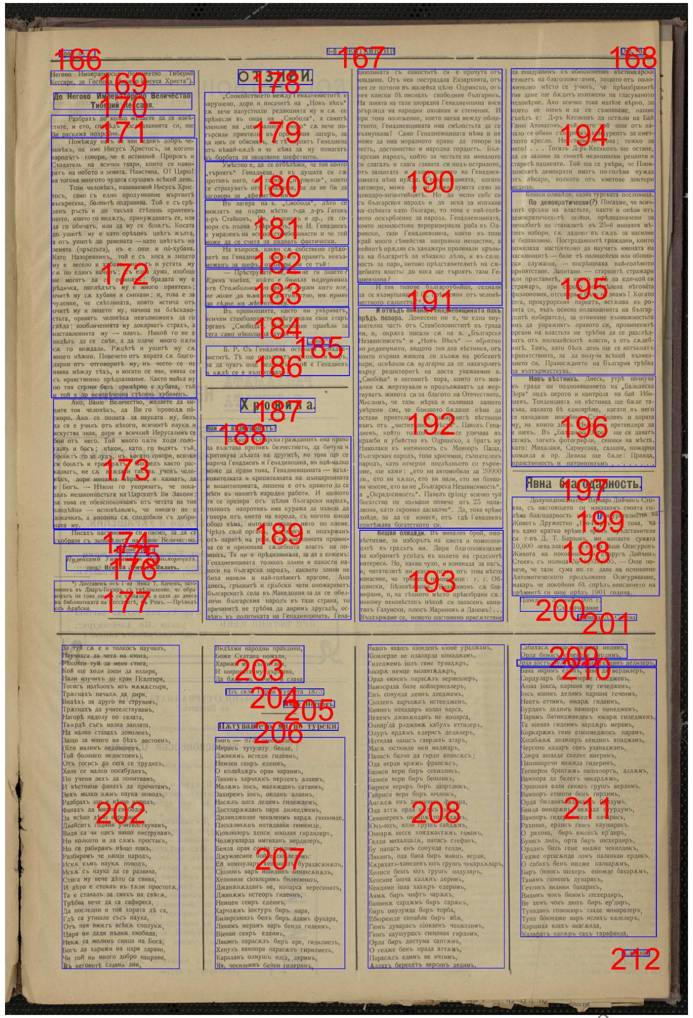
Да погледнемъ сега на нашата изложба отъ друга страна. Всѣкий народъ се гордѣе съ това, което е направилъ полезно за себе си и за другитѣ народи.

Training snippets (*on the left*) and the recognized text (*on the right*).

Extraction of Reading Orders

- Our algorithm works with **blocks of text.**
- We extracted the initial blocks with the pdfminer library.
- We view them determined by 4 numbers: (x1, x2, y1, y2) — **2D coordinates**
  - bottom left corner — (x1, y1)
  - top right corner — (x2, y2)



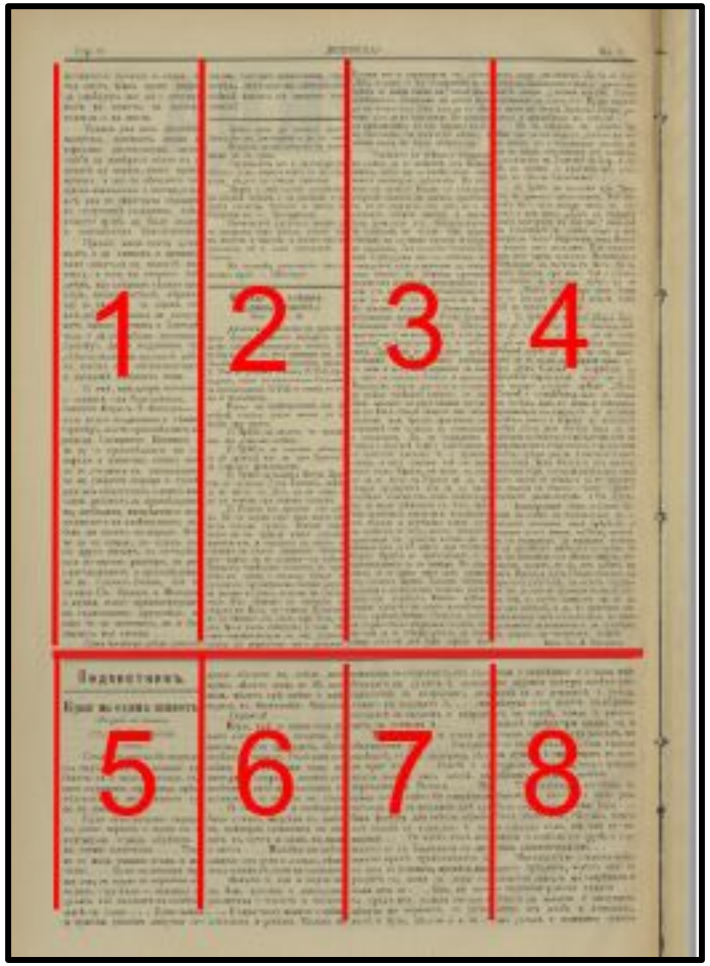


Reading order algorithm

- As a first step, we find the **subpages of the page**.
- For this purpose, we find the **horizontal gaps** that expand from the left side to the right side of the page.

Reading order algorithm

- After the identification of the subpages, our next goal is to find the **subpage columns**.
- The boundaries of the columns are recognized as **vertical gaps**.



Reading order algorithm

- In analogy to the vertical separators that define columns, there are also **partial horizontal separators** that <u>change the reading</u> order between the columns.

Благодарность.

Reading order algorithm – Sort

After these steps, we employ a **lexicographical sort** on all blocks by:

1. page number
2. subpage number
3. the numbers assigned by each of the partial horizontal separators
4. column number
5. top y coordinate
6. left x coordinate

Evaluation

- The performance of the algorithm depends on its parameters. The borderline cases of too large or too small values suggest automatic fine-tuning over the parameter values of the algorithm.
- In order to find the best parameters, we need a **gold set** of examples and a method of evaluation over the constructed reading order.
- As a metric for the evaluation of the reading order, we defined a generalized version of the **Levenshtein Edit Distance over blocks** with respect to the changing operations.

Evaluation

- Edit distance between the manually annotated reading order and:
  - the off-the-shelf library extraction (**Baseline**)
  - our page processing algorithm with initial intuitive set of parameters (**Initial**)
  - our page processing algorithm with optimized parameters over the gold dataset (**Optimized**)

Baseline	Initial	Optimized
2874	1925	1702

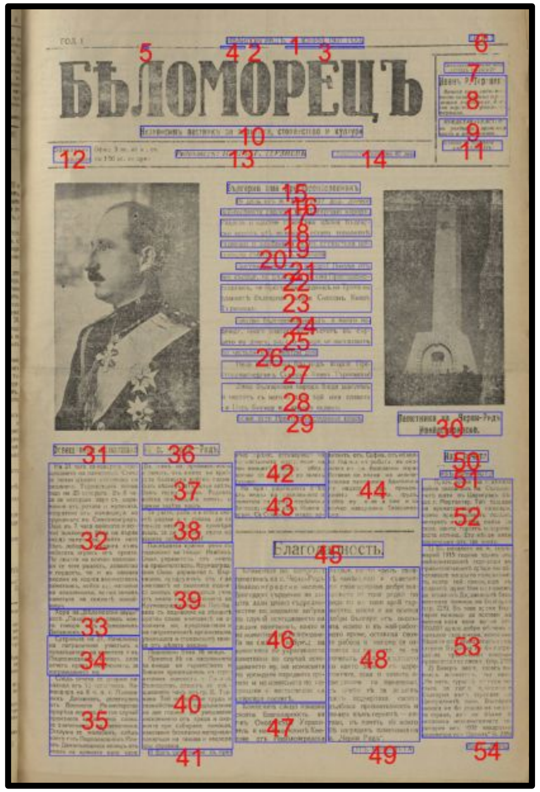
Evaluation

- Optimal values of the reading order parameters of the algorithm according to an extensive grid search over the gold dataset.

Parameters	**Initial**	**Optimized**
x_step	5	5
x_tolerance	10	12
y_tolerance	20	20
subpage_gap_threshold	10	9
partial_gap_threshold	20	26
min_column_page_ratio	0.60	0.55
min_column_width	100	20
Edit distance	1925	1702

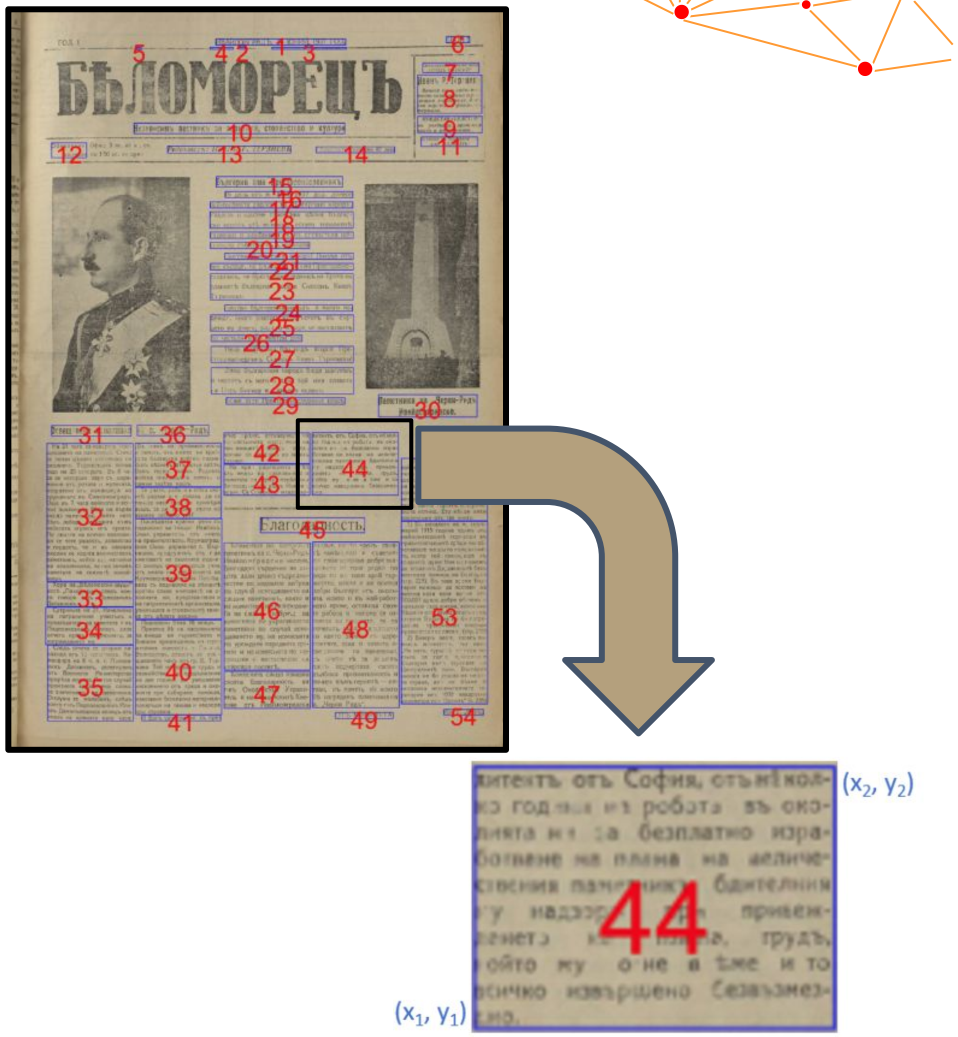
Manual Correction of the Extracted Reading Order

- The algorithm for assigning the order of blocks mostly reduces manual work. Still, there remain some problems:
    - Noise blocks or whole columns;
    - Blocks covering more than one column;
    - Wrong block order.
- Thus, if we want to produce an extraction of high quality, the manual correction is inevitable.



Manual Correction of the Extracted Reading Order

- At this stage of processing, our goal is the manual correction to be performed **on the level of blocks**.
- This would minimize the need to read all the internal text.



Manual Correction of the Extracted Reading Order

- For **visualizing and correcting**, we created an application that loads the XML file representing the blocks and the reading order as well as the corresponding PDF for presentation of the page with visualized blocks and their numbers.
- The application implements 3 commands for editing the blocks:
    - **Classification of blocks** as normal, meta, and noise.
    - Correction of the block **coordinates**.
    - **Editing the block order** by selecting the blocks to swap.

Conclusion and Future Work

- The approach we presented here **reduces the number of errors** in the reading order.
- The amount of newspaper pages to be processed – **over 150 000 pages** in more than 300 titles, makes the manual inspection a tremendous task.
- We plan to **modify and fine-tune a pre-trained encoder transformer model** to take as input subwords together with their 2D positions, and to output a text in the correct order.
- Another future task is to **add annotation** of articles and metadata to the gold dataset to fine-tune the encoder article segmentation and classification models.
- In addition, spelling translation and correction will also be performed with language models.