# DuoSearch: A Novel Search Engine for Bulgarian Historical Documents

Angel Beshirov, Suzan Hadzhieva, Ivan Koychev, and Milena Dobreva

Faculty of Mathematics and Informatics, Sofia University “St. Kliment Ohridski”, Bulgaria

## The problem

- In Bulgaria, there is already a collection of digitised historical newspapers, but access to it by the end-users is cumbersome.
- The two main issues are errors introduced during OCR process and the mixture of orthographic conventions.
- We are applying a novel approach that builds upon the automated techniques for post-OCR text correction in combination with spelling conversion.
- Our search engine was used with a subset of the historical newspaper collection from The National Library "Ivan Vazov" (NLIV) in Plovdiv for a case study.
- The purpose of our research was to build a prototype search engine which addresses the two issues mentioned above and be extendable for other languages as well.
- We have provided a live demo and open-sourced our code.

Live demo

Source code

## System Design

- The system uses the three-tier architecture.
- To tackle the linguistic variance issue, the search API component has a converter which transforms the text entered by the user into the historical spelling.
- The processor component does the data preprocessing, which includes correcting mistakes from the post-OCRed text.

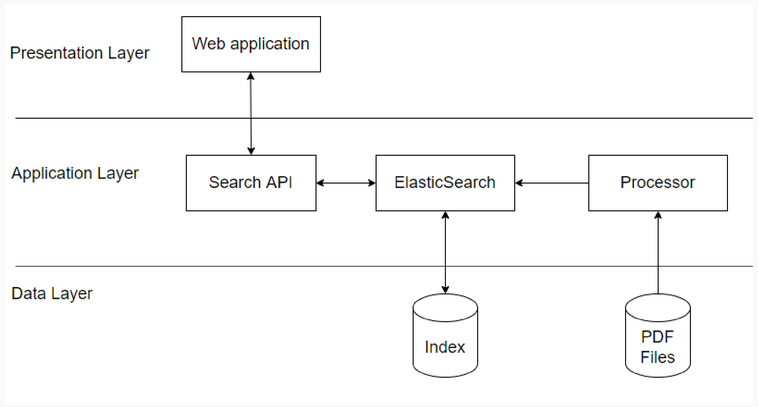


Fig. 1: System architecture

- The post-OCR text correction is separated into two inter-dependant tasks: error detection and error correction.
- For error detection we have used pretrained multilingual BERT together with a Convolutional Neural Network.
- For error correction we have used character-level sequence to sequence model with a dictionary in the old orthographic convention.
- The search engine supports two types of search: regular search and extended search.

## Evaluation

- For evaluation we have used the Bulgarian dataset provided by the organizers of the ICDAR 2019 competition.
- For evaluation metrics of our text correction model we have used F-score and % of improvement.

| Model | F-score | % of improvement |
|---|---|---|
| Clova AI | 0.77 | 9% |
| DuoSearch | 0.79 | 18.7% |

Tab. 1: Evaluation results

## Conclusion and Future Work

- The search engine prototype combines various technologies to allow for fast searching across a collection of historical newspapers.
- It has been acknowledged by Europeana as an example of successful partnership between universities and libraries.
- In future, we will work on improving the text correction part for documents containing a mixture of orthographic conventions.