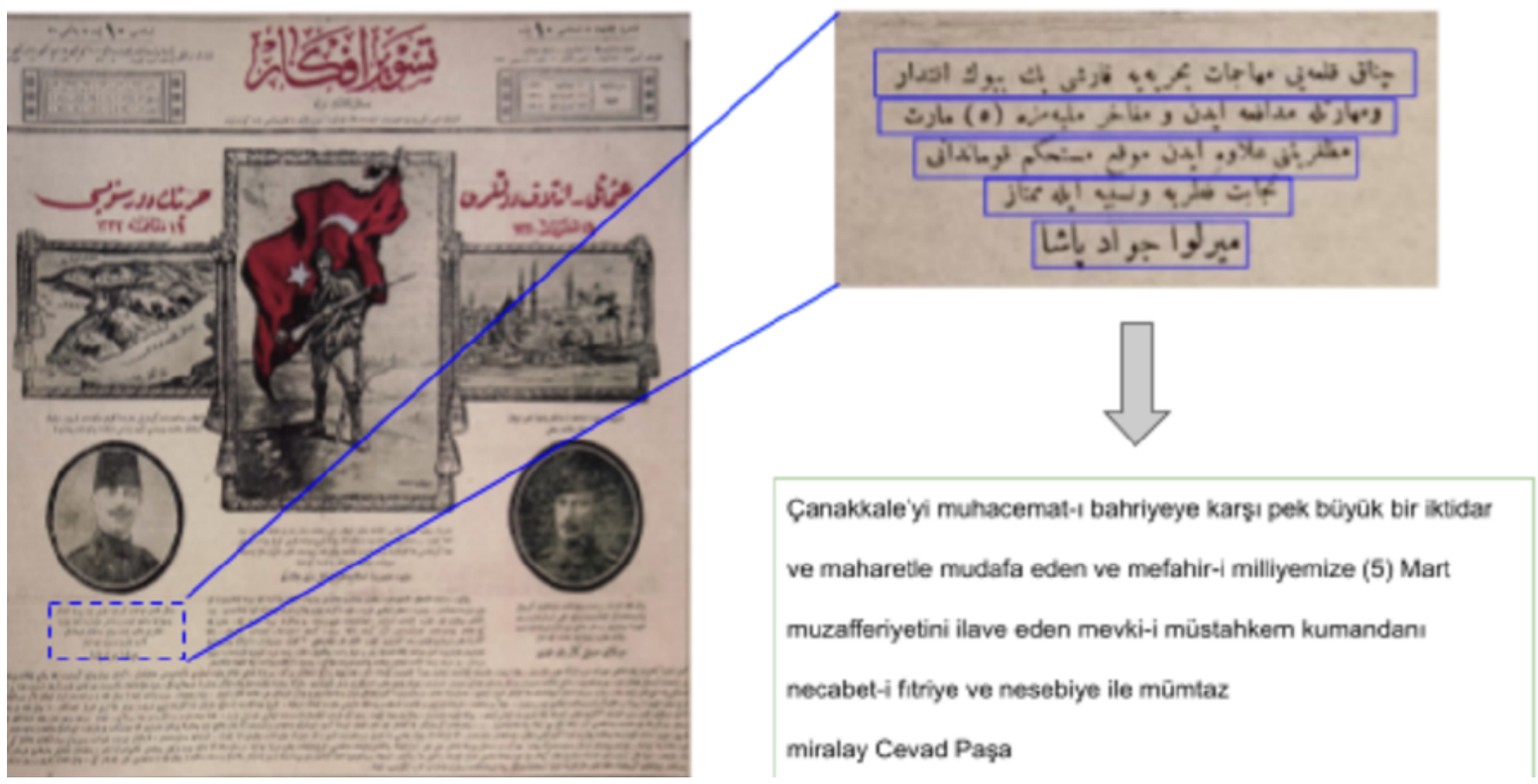
AKİS: TRANSCRIPTION OF OTTOMAN-TURKISH TEXTS USING DEEP-LEARNING

YANIKOĞLU, B., KURU, M., BILGIN TAŞDEMIR, E.F., AKCAN, A., KIZILIRMAK, F., ÖNCEL, F.

INTRODUCTION

The speed of the digitization process of Ottoman-Turkish sources is increasing every passing day and dealing with this gigantic corpus has now begun to require much more than individual human labor. Given these realities, considering the deep-learning era of AI-assisted OCR-based technologies, the automated transcription of the printed and handwritten materials will provide us with a path to access big data for the prospective research on the field of Ottoman Studies. For these very reasons, the AKİS project is initiated by computer scientists and historians collaboratively and aims to eliminate the manual transcription and to develop deep learning-based document segmentation and handwriting recognition technologies towards the automatic transcription of Ottoman printed texts and handwritten manuscripts.



WORKFLOW & METHODOLOGY:

Image Pre-Processing and Line Segmentation:

- ***Skew correction, noise removal, cleaning***
- ***Segmentation the page image into text columns and lines***

Annotation:

- ***Each training sample is input to the recognition system along with its annotation which is the ground truth for that sample.***

Training the Recognition Engine:

- ***The pre-processed and transcribed lines are used to train a deep neural network which is a Long Short-Term Memory (LSTM) network.***

EVALUATION:

- ***The results from the feasibility study we carried out to validate our approach***
- ***The dataset contains around 700 pages, 14.000 lines, 110.000 words and 30.000 unique words.***
- ***We used 70% of the data for training, 20% for validation and 10% for testing.***
- ***The system, which is trained with almost 10.000 text line images, has obtained 7.17% CER on the test. In other words, almost 93% of the characters in the transcription are recognized correctly.***

Uyuyan hayatını kendi tabiri üzere bin türlü ihtirasın, bin türlü çerağ

DATASETS:

The criteria for selecting a document are based on

- ***font and the writing style***
  - *images of printed texts with Nesiḫ (Naskh) font, and then images of handwritten texts with a clean background and that were written in the same writing style will be used for training the system.*
- ***content***
  - *Manuscripts prepared before the printing press era: Manuscripts from the first period of printed works in the dataset; the manuscripts has a wide range from the 15th to 20th centuries;*
  - *The first period of the printing press era: Texts from the years 1727-1848, mainly the printed versions of the manuscripts produced in the previous centuries;*
  - *Late period of the printing press era: Products of vibrant print media in the middle of the 19th century till the Language Reform in 1928.*
- ***layouts***
  - *different page layouts*
- ***image resolution***
  - *the image resolution should be at least 300 dpi*

*We owe many thanks to our colleagues Esmanur Sönmez, Selami Doğan Akansu, Sinem Azade, Zeynep Tandoğan