Nr. 281 שנת העשרים ואחת

מחירה

הצפירה

NO TABULA RASA

Digitizing Historical Newspapers Here and Now

DATECH 2019 -Digital Access to Textual Cultural Heritage

Варшава, 21 Декабря (2 Января) 1894/5 г. Warszawa, dnia 21 Grudnia (2 Stycznia) 1894/5 r. ווארשא, יום ד' ו' טבת התרנ"ה

זריחת החמה ש' 8 מ' 6 · שקיעת החמה ש' 3 מינ' 52 · אורך היום ש' 7 מינוט 46 · סוף זמן בין השמשות ש' 4 מינ' 39.

אדרעססע: רעדאקציה הצפירה קרולעווסקא № 49 ווארשא.

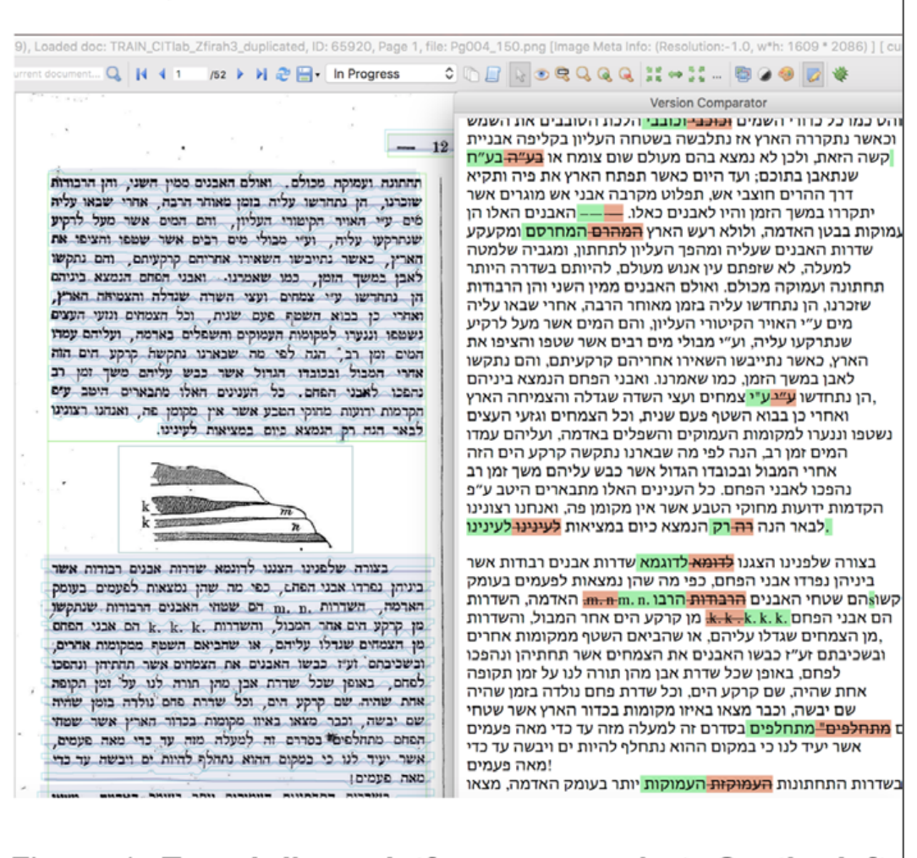
Introduction

We share the workflow, the script and the data involved in the process of improving a corpus without losing the fruits of previous work to analyse document structure. The amelioration of text recognition is conducted within a structured corpus of a Hebrew Language 19th Newspaper. This includes conversion of Abbyy OCR text-region output and integration of Olive Software Preservation Markup Language Schema (PRXML) into the PAGE XML format used by Transkribus; adaptation of Transkribus' line detection to our RTL newspapers; XML-TEI export to external tools and the use of Transkribus API to automate the workflow. The project will not only enable thorough digital research of a fascinating historical corpus of an important newspaper, the first daily in the Hebrew language; it also serves as a proof-of-concept study for any future endeavour to salvage 315 other titles, entailing over two million pages, in the Jewish Historical Press collection JPRESS, of Tel Aviv University and the National Library of Israel. The workflow can also be adopted by hundreds of digital collections worldwide which are in similar condition. Finally, the discussion will draw lessons regarding the challenges for work along developing methods, tools and platforms

Hebrew Text Recognition training and interpretation

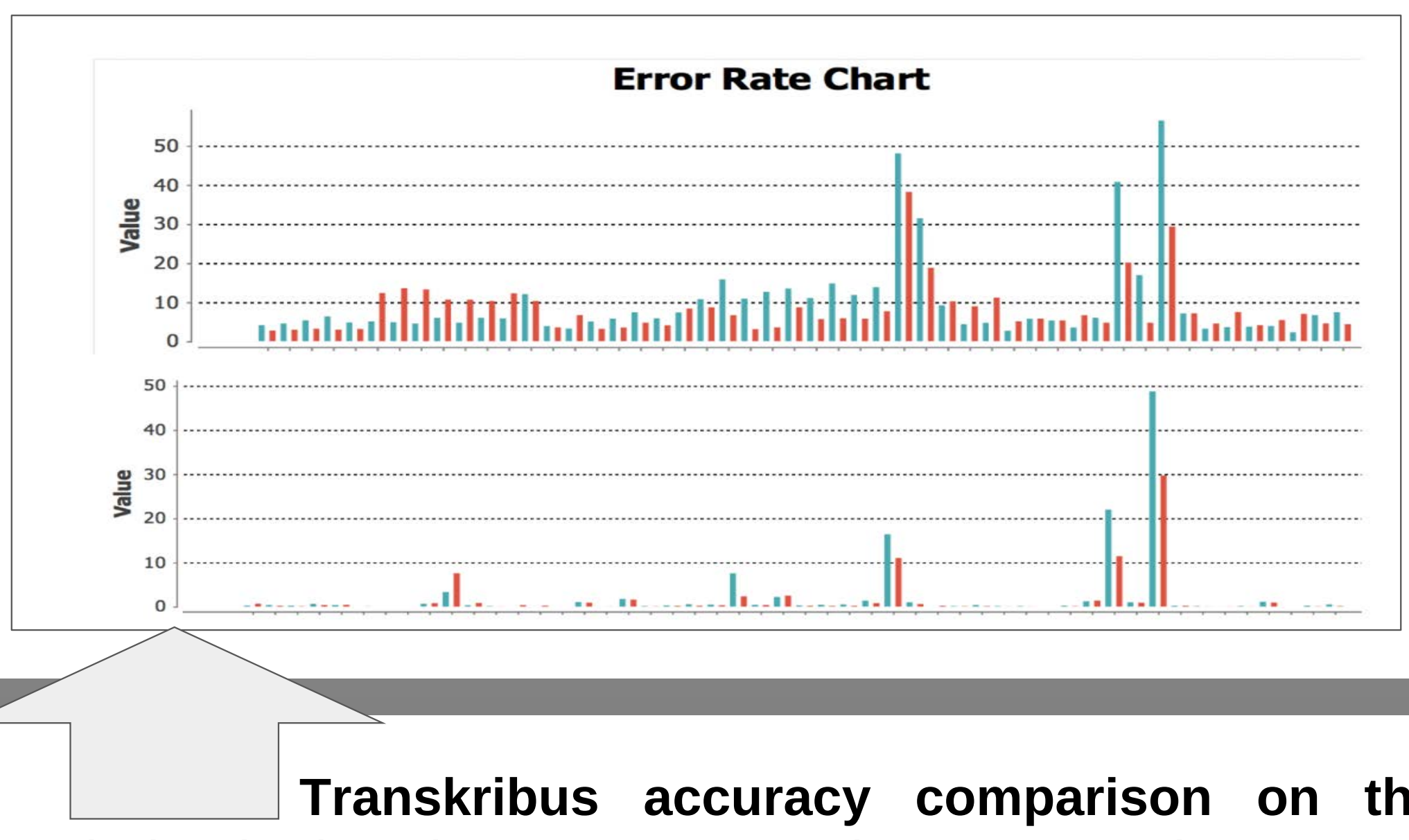
Transkribus was used to create a ground-truth corpus of 50 pages and train a model for historical Hebrew print. The ground truth was based on issues from the entire publication span of 1874, as well as issues from 1875 and 1887. The corpus has varying scan quality, and includes, in addition to the 'Meruba'(square) Hebrew script, articles in Rashi script which was used for the halachic sections, (discussions of religious law). The resulting model, titled "HaZefira3" yielded an average accuracy of 96.6% on the trained set, and 96.7% on the test set. Training with Transkribus' HTR+ model showed significant improvement, as shown in the figure below:

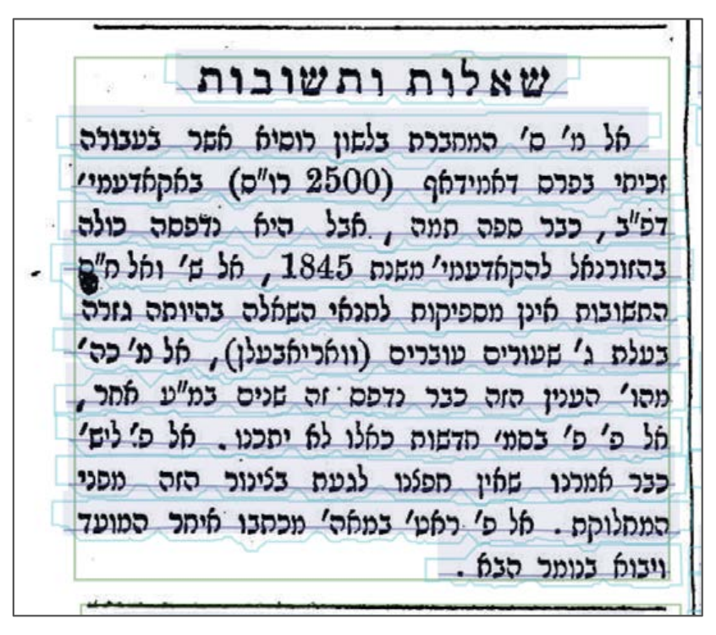
"HaZefira3" yielded an average accuracy of 96.6% on the trained set, and 96.7% on the test set.



9), Loaded doc: TRAIN_CITlab_Zfirah3_duplicated, ID: 65920, Page 1, file: Pg004_150.png [Image Meta Info: (Resolution:-1.0, w*h: 1609 * 2086)] [cu

Figure 1: **Transkribus platform screenshot. On the left, the analysed and OCR'd page; on the right, a view of the HAZFIRAH3 model evaluation of the same page.**





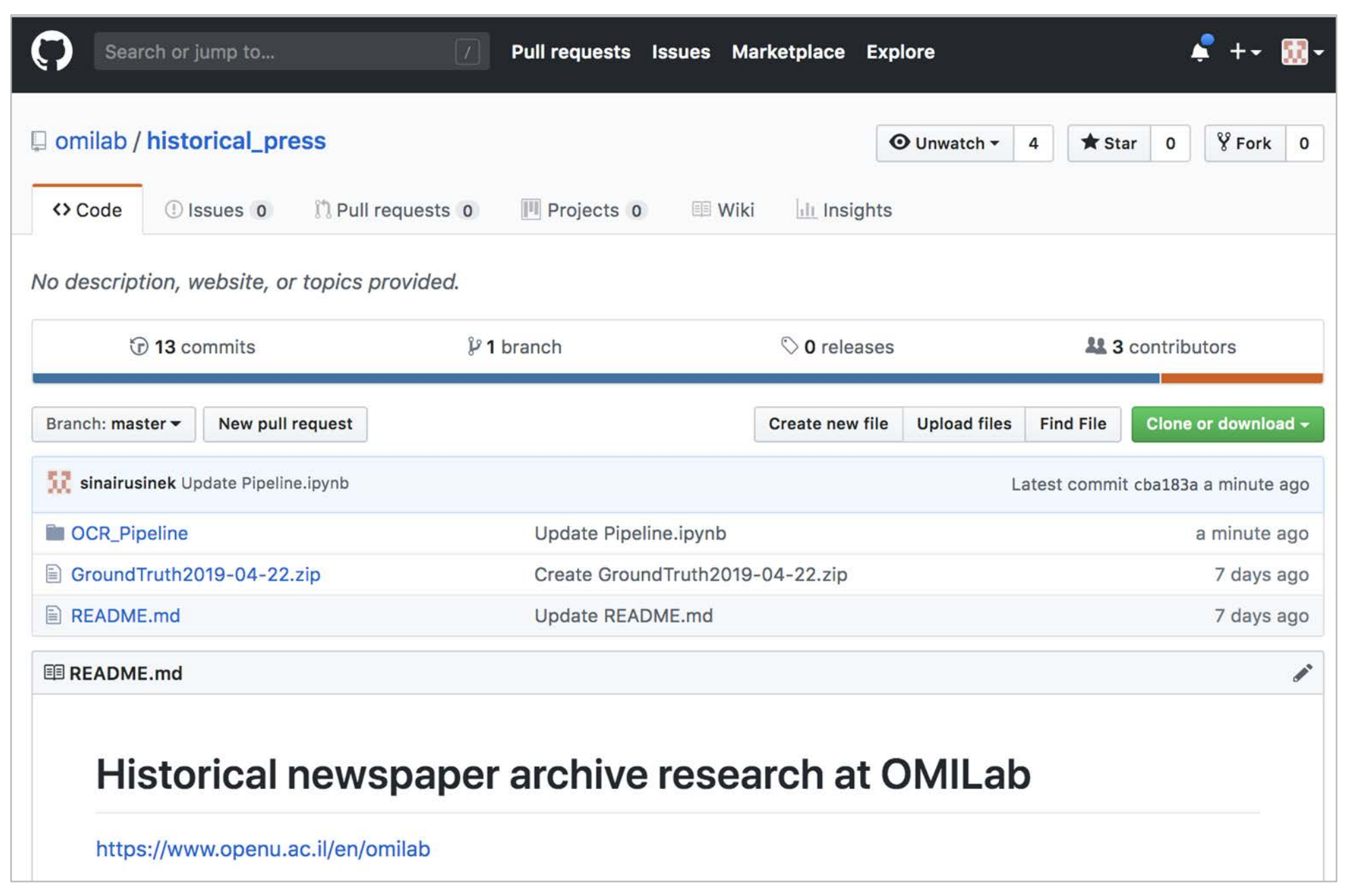
שאלות ותשובות

אל מ' ס' המחברת בלשון רוסיא אשר בעבודה
זכיתי בפרס דאמידאף (2500 רו"ס) באקאדעמיי'
דפ"ב, כבר ספה תמה, אבל היא נדפסה כולה
בהזורנאל להקאדעמיי' משנת 1845, אל ט' ואל ח"ס
התשובות אינן מספיקות לתנאי השאלה בהיותה גזרה
בעלת ג' שעורים עוברים (וואריאבעלן), אל מ' כה'
מהו' הענין הזה כבר נדפס זה שנים במ"ע אחר,
אל פ' פ' בסמי' הדמות כאלו לא יתכנו. אל פ' ליס'
כבר אמרנו שאין חפצנו לגעת בלינור הזה מפני
המחלוקת. אל פ' ראט' במאה' מכתבו איחר המועד
ויבוא בנומר הבא.

Similar success was achieved for the 'RASHI' script, a semi-cursive typeface for the Hebrew alphabet, used in portions of the newspaper issues.

The Pipeline

Available as Jupyter notebook on Github, a modular pipeline to automate the migration process from the legacy PRXML to PAGE.xml, preserving the structural information.



omilab / historical_press

No description, website, or topics provided.

13 commits · 1 branch · 0 releases · 3 contributors

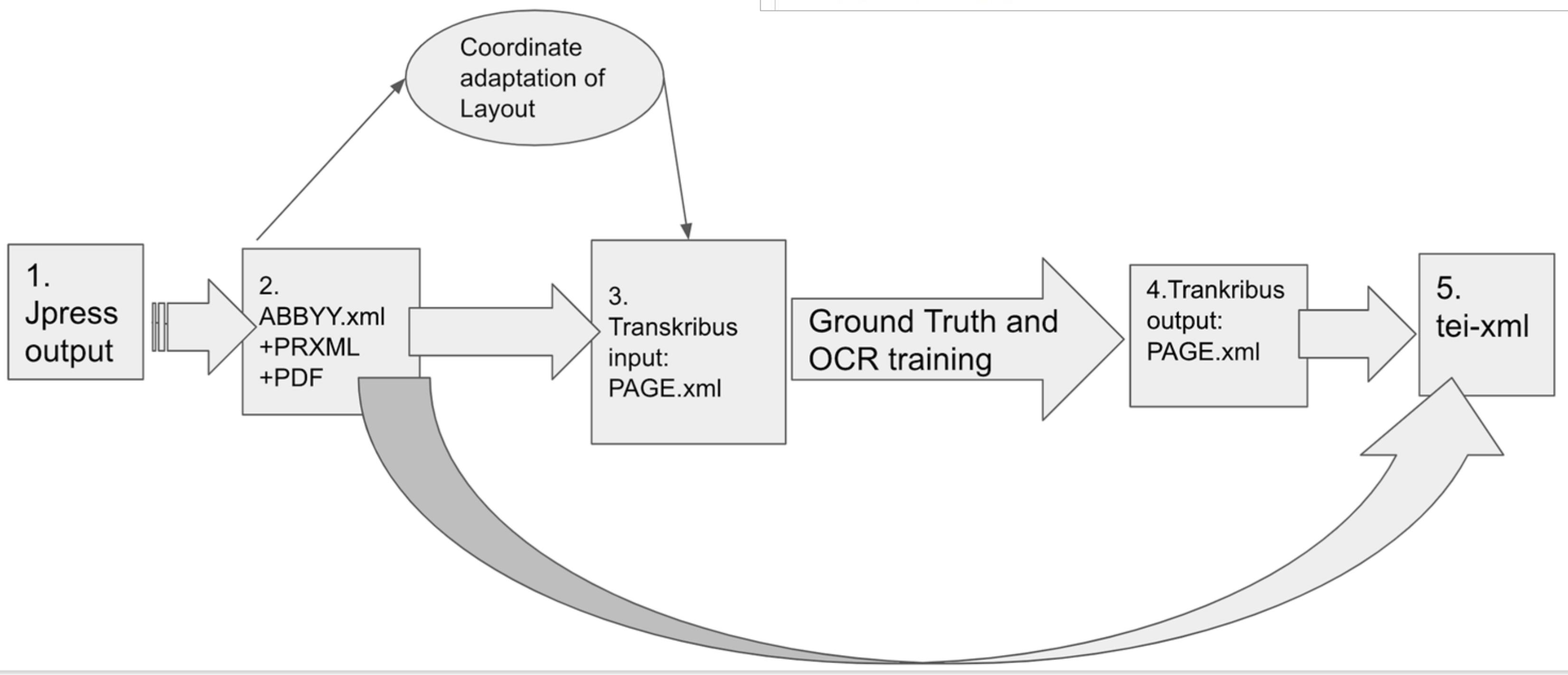
sinairusinek Update Pipeline.ipynb — Latest commit cba183a a minute ago

OCR_Pipeline	Update Pipeline.ipynb	a minute ago
GroundTruth2019-04-22.zip	Create GroundTruth2019-04-22.zip	7 days ago
README.md	Update README.md	7 days ago

README.md

Historical newspaper archive research at OMILab

https://www.openu.ac.il/en/omilab



Coordinate adaptation of Layout

1. Jpress output

2. ABBYY.xml +PRXML +PDF

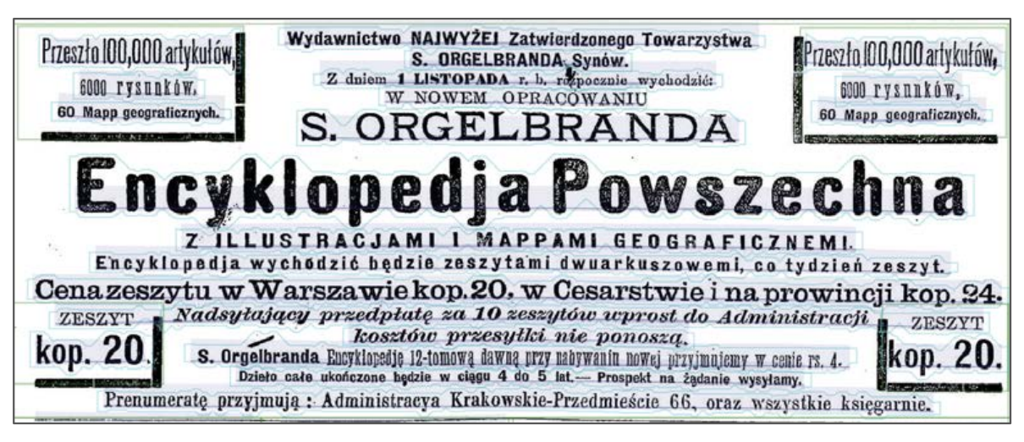
3. Transkribus input: PAGE.xml

Ground Truth and OCR training

4.Trankribus output: PAGE.xml

5. tei-xml

Transkribus accuracy comparison on the 50 ground truth pages are revealing; Each line in the figure represents the error rate in a page. **Blue represents WER (word error rate), red represents CER (character error rate). Above: HTR based model. Below: HTR+ based model.** Both graphs shows outliers with rather high error rates: these are caused by a back page from 1887 which contained long advertisements in Polish, rather than the Hebrew language (It will not be included in the final ground truth), and pages of index for the articles in 1974, containing short lines with mainly names, titles and page numbers. These were more challenging for the automatic reading because of the large number of digits,



Przeszło 100,000 artykułów, 6000 rysunków, 60 Mapp geograficznych.

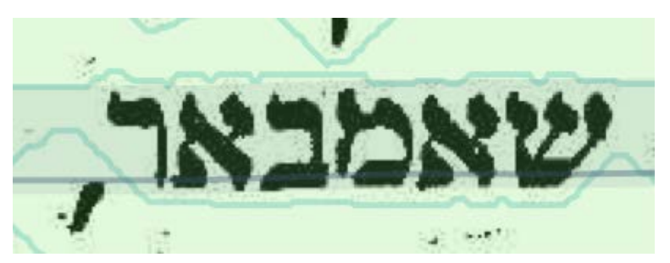
Wydawnictwo NAJWYŻEJ Zatwierdzonego Towarzystwa
S. ORGELBRANDA Synów.
Z dniem 1 LISTOPADA r. b. rozpocznie wychodzić:
W NOWEM OPRACOWANIU
S. ORGELBRANDA
Encyklopedja Powszechna
Z ILLUSTRACJAMI I MAPPAMI GEOGRAFICZNEMI.
Encyklopedja wychodzić będzie zeszytami dwuarkuszowemi, co tydzień zeszyt.
Cena zeszytu w Warszawie kop. 20, w Cesarstwie i na prowincji kop. 24.
Nadsyłający przedpłatę za 10 zeszytów wprost do Administracji kosztów przesyłki nie ponoszą.
S. Orgelbranda Encyklopedję 12-tomową dawną przy nabywaniu nowej przyjmujemy w cenie rs. 4.
Dzieło całe ukończone będzie w ciągu 4 do 5 lat.— Prospekt na żądanie wysyłamy.
Prenumeratę przyjmują: Administracya Krakowskie-Przedmieście 66, oraz wszystkie księgarnie.

ZESZYT kop. 20.

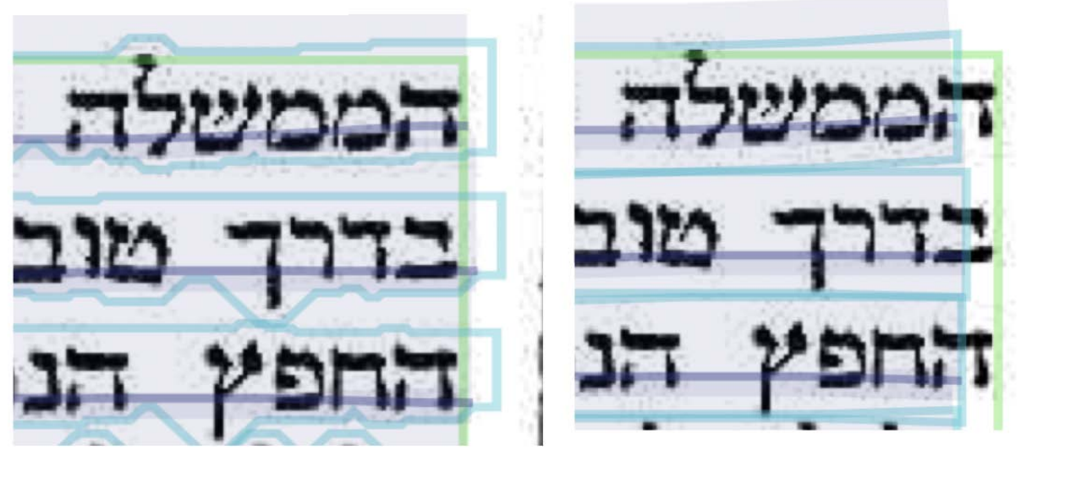
Przeszło 100,000 artykułów, 6000 rysunków, 60 Mapp geograficznych.

ZESZYT kop. 20.

The more interesting phenomenon, however, is the relation of WER to CER which changes in sections of the corpus, and can be revealed in the upper figure. The reason was found to be related to line detection: in pages 6-13 of the corpus, which came from a 1874 issue, the red CER is higher than the blue OCR, because most of the errors were related to commas and dots that fell outside the line. This would raise CER but not harm word accuracy



In pages 22 to 29, however, WER is significantly higher than CER. These pages are taken from an 1887 issue, containing 3 columns per page, as opposed to two columns in the older issues. The additional column impacted the accuracy of baseline end detection and letters were either cut entirely or partially at the beginnings and ends of line, so here again the OCR error were only an indicator of a layout analysis failure. As on the right, having manually fixed in the Ground truth set, and a new line detection model was trained. The accuracy comparison above shows the impact on the new model.



הממשלה
בדרך טוב
החפץ הנ
הממשלה
בדרך טוב
החפץ הנ

Figure 4: **Transkribus line indentation problem, seen here on the line beginning (right). on the left, the extended baseline version.**

Contact us!

Sinai Rusinek
Haifa University and The Open University, Israel
sinai.rusinek@mail.huji.ac.il

Nurit Greidinger
The Open University Israel
nuritgr@post.bgu.ac.il

Адресъ: Редакція,, Гацефира", Варшава, Королевская 49 | Adres: Redakcya ,,Hazefira" Warszawa, Królewska 49

Дозволено Цензурою, Варшава 21 Декабря 1894 г. Въ Типографіи М. Левинскаго, Варшава. Краковское-Предмѣстье 40.

Редакторы-Издатели: З. Слонимскій и Н. Соколовъ.