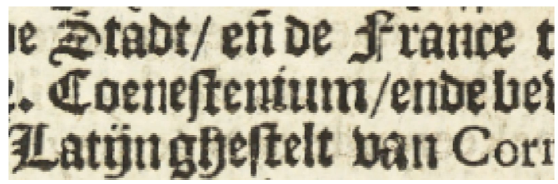
Improving the recognition of Dutch Gothic machine print, at four levels in the processing pipeline, in four days

Lambert Schomaker, Mahya Ameryan, Mirjam Cuper, Koen Dercksen, Jerry Guo, Rutger van Koert, Adriënne Mendrik, Konstantin Todorov, & Xue Wang

Report on an *NWO/ICT with industry project* (5-day sabbatical/hackathon) focusing on OCR problems at the Dutch National Library (KB) and the Huygens Institute, January 20-24, 2020, Lorentz Center, University of Leiden


ie Stadt/ eñ de France t
.. Coeneſtenium/ende be
Latijn gheſtelt van Corr

1. Problem statement

Libraries and archives are struggling with optical character recognition (OCR) of old machine-print fonts such as Gothic or 'fraktur'. This font was used in many important historical printed collections such as administrative texts and the then (17th century) newly invented 'newspapers' with interesting and detailed reports on important developments and events. When applying current state of the art OCR tools or sending the scanned images to large well-known companies that provide OCR services, the returned results are still quite disappointing. Problems are observed at all levels in the processing pipeline: binarisation suffering from ink bleed-through, layout analysis suffering from deviating page designs, marginalia and graphics, character recognition suffering from lack of pertinent font examples and font variation (Roman/Gothic) in a document and, finally, linguistic post processing suffering from an utter lack of encoded digital text corpora of suitable size. Actually, the OCR process is often intended to arrive at such corpora in the first place.

2. Approach

A team was formed to approach these problems in four days, with a fifth day for reporting (other teams were working on other industrial problems at the Lorentz Center, this week). The team decided to address problems at all levels in the processing pipeline.

3. Team composition

Industry: Dutch National Library (KB) & KNAW Humanities Cluster (Huc)
- Mirjam Cuper and Rutger van Koert

Academia: - prof. dr. Lambert Schomaker (team leader, University of Groningen, AI/ML)
In alphabetic order: Mahya Ameryan (AI, University of Groningen), Koen Dercksen (Radboud University), Jerry Guo (Technical university, Delft), Konstantin Todorov (ILLC, University of Amsterdam), dr. Adriënne Mendrik (e-Science center, The Netherlands), Xue Wang (CS,

Leiden University)

4. Tasks

The processing pipeline consists of (1) image preprocessing, (2) layout analysis and segmentation into meaningful text objects (lines, words, characters), (3) text recognition and (4) linguistic post processing.

Adriënne Mendrik focused on modeling the overall process in order to arrive at a specification of a modular work flow including performance evaluation, using her expertise on similar processing pipelines encountered at the Dutch e-Science center, in medical imaging applications.

Preprocessing: Jerry Guo (binarisation)
Layout analysis: Xue Wang (word segmentation)
Text recognition: Mahya Ameryan (characters), Lambert Schomaker (words)
Post processing: Koen Dercksen, Konstantin Todorov

5. Data & research questions

The first day of the week was used to delve into the problem. Schomaker gave a crash course at the whiteboard on OCR to the participants who were from different research fields and literature hints were disseminated. The problem owners presented a host of data sets and problems. The post-processing group discovered an SQL database with recognized text and wanted to relate it back to the original images. In itself, this appeared not to be possible because the work flow in the application domain is more focused on the encoded digital text than on image (pixel) related information. However, inspection of the Alto .xml files from the commercial OCR company that provided recognition results, it appeared that the bounding boxes (ROIs) of recognized strings were present in the .xml. This data set proved to be usable by all participants. It consists of a newspapers collection in Dutch Gothic script (1662 – 1795) from the Meertens Institute (Amsterdam), 15172 scans in Jpeg2000 format, which is common in libraries and archives, but not in science. The scans were converted to Jpeg by van Koert, who also provided a server with GPUs. Collaboration tools were Google Drive, Slack and Surf storage. The ensuing research questions (RQ) are:

1. Can we improve the recognition results by improving the binarisation of the scans?
2. Can we improve the detection of words, because the commercial OCR providers often produce very long strings in the recognized output, with no blanks between words
3. Given the problems at the character level, wouldn't it be better to train at the level of words, and how fast can we collect word labels from a tabula rasa setting, using high-performance computing and a human in the loop (the Monk approach)
4. Contrarily, can we use the suboptimal recognition results from a commercial provider (i.e., without human-validated ground truth) to train our own recognizers from scratch and allow many recognizers to be applied to the same data?
5. What are the effects of using state-of-the art text transduction tools to improve the quality of the OCR output to something that is closer to the intended language?
6. How does the performance evaluation work, at the different four stages in the processing pipeline?

6. Methods

RQ1. Binarisation – The approach is to apply a number of traditional and deep-learned binarisation methods to a selected subset of scans from the Meertens newspaper data. Because no ground truth on the pixel intensities (ink/background) exist, as in the international DIBCO competition, another performance indicator is used, i.e., the final recognition performance of a common free OCR tool, Tesseract, on the data, which is used as the measuring device.

RQ2. Word segmentation – The HUC (problem owner) already had experience with using ARU-net (a U-net CNN, i.e. a deep-learning method) for baseline estimation. Because the Alto .xml output contains the ROI on the position of blanks via the '<SP>' tag, a new network can be trained, end to end, combining the line segmentation and word segmentation in an innovative single pixel-to-pixel transform.

RQ3. Full-word recognition – The existing Monk e-Science service was used to bootstrap a label collection process, starting from the tabula rasa condition), i.e., no labels. Scans were segmented into lines automatically and oversegmented into word candidates. Users start to label some correctly segmented words and an immediate process of data mining starts in the background using nearest-neighbour matching (upon the first label) and nearest-mean matching (with Nlabels > 1), in a high-performance computing setup. This is done using an alternation between recognition and retrieval, yielding ranked hit lists that can easily be labeled: the 'Fahrkunst' principle, Schomaker & van Oosten (2014).

RQ4. Fresh, end-to-end training of a CNN/LSTM using the suboptimal recognition results of a commercial OCR provider as the training set. The OCR output is pruned to contain word-like strings, starting with a letter. A small ensemble of five classifiers using plurality voting and tie resolution is used. The performance evaluation is case insensitive.

RQ5. A current state-of-the art approach to language translation is to use word embeddings (cf. Latent semantic indexing and the later word2vec approach) followed by an LSTM encoder and an LSTM decoder. For word embeddings, we use BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (2018), a pretrained network that is fine tuned on the text to be expected in the Meertens Newspapers corpus.

RQ6. How to provide a modular pipeline and work flow, such that users in the application domain can quickly evaluate different variants of processing? Ideally, tests are performed on the grounds of a neutral party, such as a national e-Science server. Is this possible in all cases, e.g. using virtualization or 'Docker' containers? In any case the overall concept for the framework can be designed in the short time frame of the current project.

7. Results

RQ1 – binarisation
Jerry Guo demonstrated that binarisation methods are less important than envisaged. The traditional methods deteriorate the OCR results. Deep-learning (Deep Otsu) does not deteriorate the recognition performance noticeably, but leaving the input images – as is – gives the OCR performance that can be reached. Tesseract was used as the ‘thermometer’.

RQ2 – word segmentation
Xue Wang showed that deep-learning based line segmentation (ARU-net based) can be enhanced by adding word-segmentation suggestions, in a pix2pix manner. The model was trained end-to-end, using recognition results (Alto.xml) of a commercial provider.

RQ3 – full-word recognition, from tabula rasa
Lambert Schomaker and Mirjam Cuper showed that the Gothic text could be ingested by the Monk system. Manual labeling with machine support yielded 319 word classes, 1760 human-labeled images in under two hours. This data can be used to evaluate different recognizers.

RQ4 – training a new recognizer on the basis of other OCR
Mahya Ameryan showed that an LSTM developed at AI@RuG could be trained end-to-end, achieving a word-recognition rate of 88%. Notably, the recognizer was trained on the output of a commercial recognizer, i.e., using word regions of interest and their OCR text as ground truth (100% can not be reached). The result is favorable and compares to a character recognition accuracy of 97.5% for a language with 5-letter words, on average.

RQ5 – linguistic post processing
Koen Dercksen and Konstantin Todorov used the famous BERT (word-embedding method by Google) system to fine tune it (post train it) using the Meertens Newspaper texts, with promising results.

RQ6 – performance-evaluation framework
Adriënne Mendrik modeled the work flow and processing pipeline in order to realize a modular, adaptable framework in which individual component variants can be inserted to assess their effect on evaluation. Performance metrics are considered within their particular usage context, e.g., transcription and retrieval require different performance metrics, and the object under consideration (characters vs words) also plays a central role in understanding the performance of an OCR system.

8. Conclusion

This was an exciting project week with tangible results, that are expected to have an impact on the ‘industry’, i.e., the KB and Humanities Cluster (HuC).
Notable is the fact that there is much more information in the recognition results of commercial providers than is currently used. Massive amounts of training data can be extracted. Apparently good recognition results can be achieved by alternative recognizers that

are trained on such imperfect data. By using user-friendly labeling tools, benchmark data sets can be developed by the institutes themselves, which puts them in a more comfortable position in dealing with OCR providers. The results present a promising picture regarding the future integration of Gothic-font documents in current search engines such as ‘Delpher’.

9. References

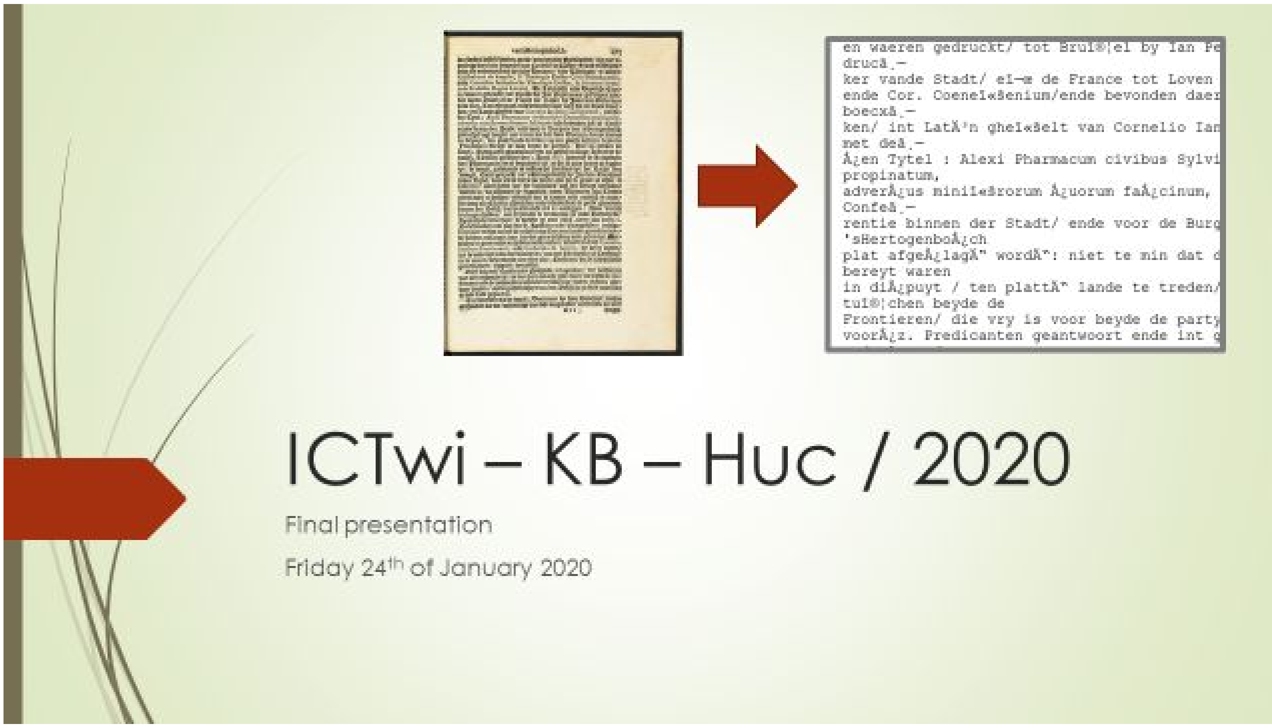
Ameryan, M. & Schomaker, L.R.B. (2019), A limited-size ensemble of homogeneous CNN/LSTMs for high-performance word classification, arXiv:1912.03223

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2018).
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv:1810.04805 [cs.CL]

Gruening, T., Leifert, G., Strauß, T., Michael, J. & Labahn, R. (2019).
A Two-Stage Method for Text Line Detection in Historical Documents, arXiv:1802.03345 [cs.CV]

van Oosten, J.-P. & Schomaker, L.R.B. (2014). Separability versus prototypicality in handwritten word-image retrieval, Pattern Recognition, 47(3), pp. 1031-1038

10. Appendix – Final slides presented on Friday (day 5)

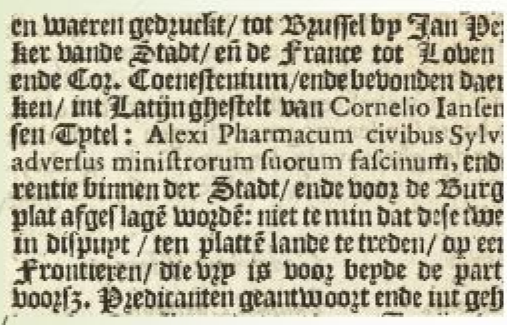


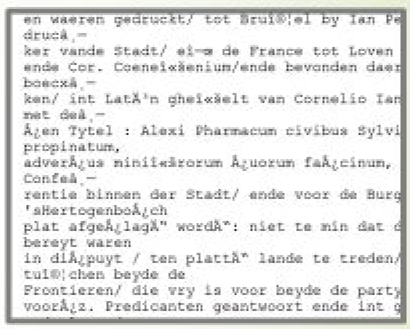
KB-Team (only the participants present at Lorentz Center during the week)

- Industry: KB & Huc KNAW
  - Mirjam Cuper (performance evaluation)
  - Rutger van Koert (tool benches image & text processing)
- Academia (team leader: prof. dr. Lambert Schomaker)
  - Mahya Ameryan (AI, RUG) – text recognition
  - Koen Dercksen (Radboud Univ.) – multimodal information processing
  - Jerry Guo (TU Delft) – computer vision
  - Konstantin Todorov (ILLC, UvA) – computational linguistics
  - dr. Adriënne Mendrik (e-Science center) – performance evaluation frameworks
  - Xue Wang (CS, Leiden Univ.) – object vision

KB What is the problem?

Mirjam Cuper
Rutger van Koert


en waeren gedruckt/ tot Brussel by Jan Pe
ker vande Stadt/ eñ de France tot Loven
ende Cor. Coenestentium/ende bevonden daer
ken/ int Latijn ghestelt van Cornelio Iansen
sen Tytel : Alexi Pharmacum civibus Sylvi
adversus ministrorum suorum fascinum, end
rentie binnen der Stadt/ ende voor de Burg
plat afgeslagẽ wordẽ: niet te min dat dese twe
in dispuyt / ten plattẽ lande te treden/ op een
Frontieren/ die vry is voor beyde de part
voorsz. Predicanten geantwoort ende int geh


```
en waeren gedruckt/ tot Brui®¦el by Ian Pe
drucâ‚¬
ker vande Stadt/ ei¬œ de France tot Loven
ende Cor. Coenei«ãenium/ende bevonden daer
boecxâ‚¬
ken/ int LatÃ³n gheï«ãelt van Cornelio Ian
met deâ‚¬
Å¿en Tytel : Alexi Pharmacum civibus Sylvi
propinatum,
adverÅ¿us minii«ãrorum Å¿uorum faÅ¿cinum,
Confeâ‚¬
rentie binnen der Stadt/ ende voor de Burg
'sHertogenboÅ¿ch
plat afgeÅ¿lagÃ" wordÃ": niet te min dat d
bereyt waren
in diÅ¿puyt / ten plattÃ" lande te treden/
tui®¦chen beyde de
Frontieren/ die vry is voor beyde de party
voorÅ¿z. Predicanten geantwoort ende int g
```

KB national library of the netherlands

KNAW Humanities Cluster

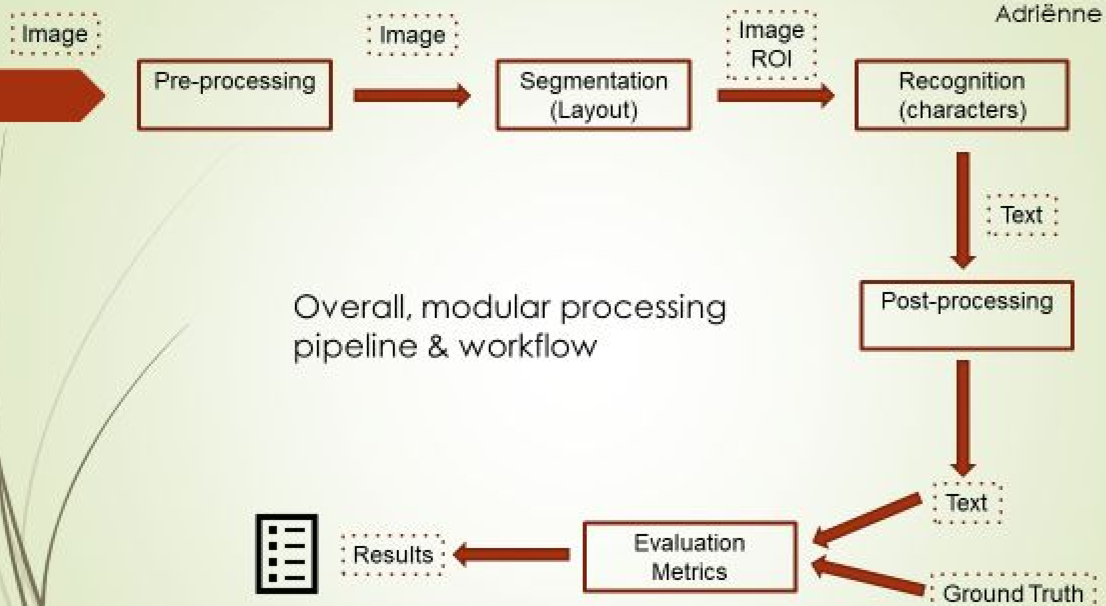
Delpher

Adriënne Mendrik

Performance-evaluation framework

Adriënne Mendrik

Overall, modular processing pipeline & workflow



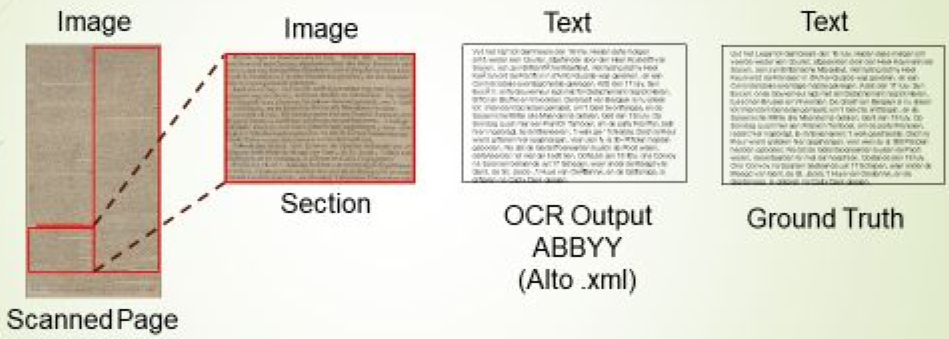
Data & environment

- Among all the data, we found a set that suited several sub teams: **the Meerten's set.** Actually Koen wanted to go from plain text back into the Alto XML, which was not possible, but then it appeared that ABBYY put the bounding boxes of recognized fragments in this metadata ☺
- Usable by Xue, Mahya, Lambert, Koen, Konstantin, Jerry (i.e.all), in each of the processing stages
- Conversion needed: Jpeg2000 (archives & libraries) is not often used in science. It needed to be converted to regular .jpg
- Rutger provided server access with GPUs, and big external disks with data
- We used: Google Drive, Slack, Surf storage
  Tensorflow & own tools, as well as some new coding (Python,Bash,C)

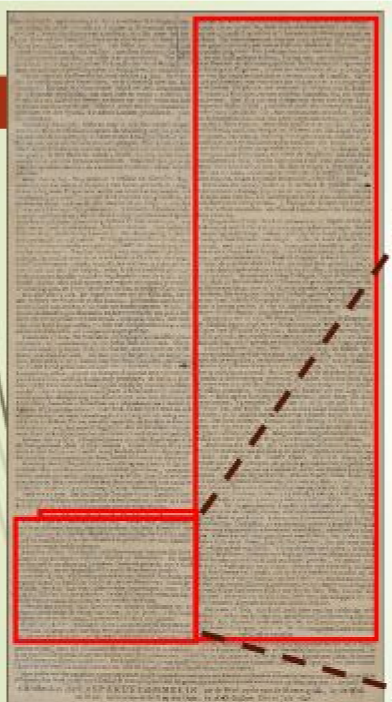
Alto.xml format snippet (Meertens collection): labels down to the word ROI in the page image

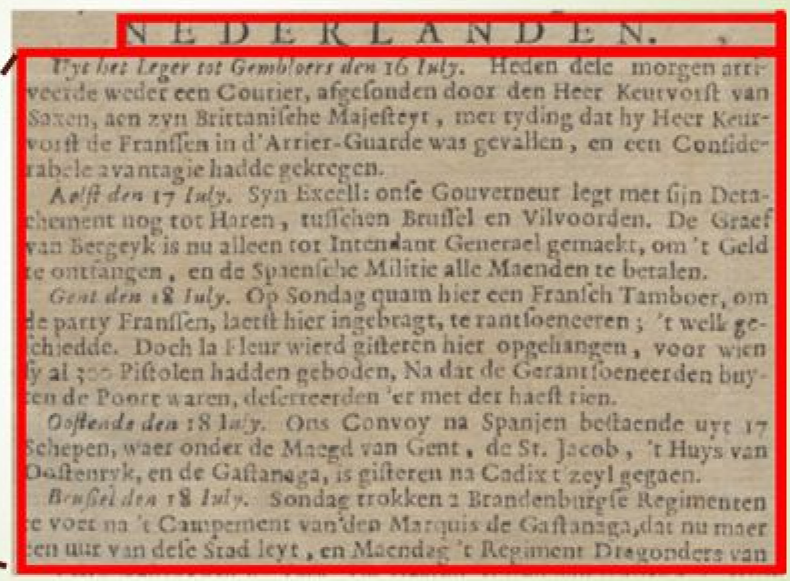
```
<TextBlock ID="P1_TB00003" WIDTH="1409" HEIGHT="100" VPOS="253" HPOS="112"
        STYLEREFS="TXT_2 PAR_LEFT">
  <TextLine ID="P1_TL00003" WIDTH="405" HEIGHT="2" VPOS="252" HPOS="112">
    <String ID="P1_ST00004" WIDTH="43" HEIGHT="2" VPOS="252" HPOS="112"
        CC="000000000" WC="0.96" CONTENT="Utrechtfe"/>
    <SP ID="P1_SP00002" WIDTH="18" VPOS="254" HPOS="155"/>
    <String ID="P1_ST00005" WIDTH="69" HEIGHT="2" VPOS="252" HPOS="172"
        CC="0000" WC="0.96" CONTENT="Vry-"/>
    <SP ID="P1_SP00003" WIDTH="18" VPOS="254" HPOS="241"/>
    <String ID="P1_ST00006" WIDTH="103" HEIGHT="2" VPOS="252" HPOS="259"
        CC="000000" WC="1.00" CONTENT="daegfe"/>
    <SP ID="P1_SP00004" WIDTH="18" VPOS="254" HPOS="361"/>
    <String ID="P1_ST00007" WIDTH="138" HEIGHT="2" VPOS="252" HPOS="379"
        CC="00677350" WC="1.00" CONTENT="Courant."/>
  </TextLine>
</TextBlock>
```

Data: Newspapers (1662 – 1795)



Original Scan: Example Segments





N E D E R L A N D E N.

Uyt het Leger tot Gembloers den 16 July. Heden dese morgen arriveerde weder een Courier, afgesonden door den Heer Keurvorst van Saxen, aen zyn Brittanische Majesteyt, met tyding dat hy Heer Keurvorst de Franssen in d'Arrier-Guarde was gevallen, en een Considerabele avantagie hadde gekregen.

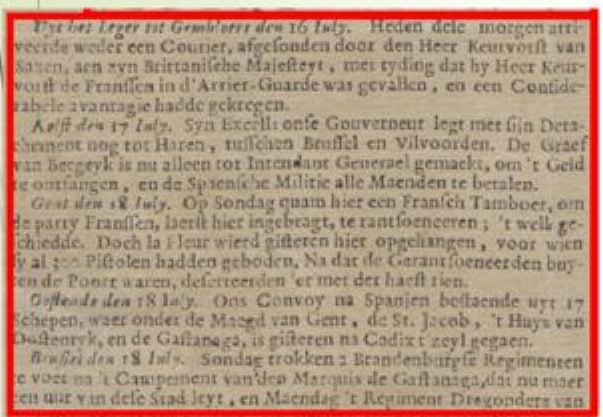
Aelst den 17 July. Syn Excell: onse Gouverneur legt met sijn Detachement nog tot Haren, tusschen Brussel en Vilvoorden. De Graef van Bergeyk is nu alleen tot Intendant Generael gemaeckt, om 't Geld te ontfangen, en de Spaensche Militie alle Maenden te betalen.

Gent den 18 July. Op Sondag quam hier een Fransch Tamboer, om de party Franssen, laetst hier ingebragt, te rantsoeneeren; 't welk geschiedde. Doch la Fleur wierd gisteren hier opgehangen, voor wien sy al 300 Pistolen hadden geboden, Na dat de Gerantsoeneerden buyten de Poort waren, deserreerden 'er met der haest tien.

Oostende den 18 July. Ons Convoy na Spanjen bestaende uyt 17 Schepen, waer onder de Maegd van Gent, de St. Jacob, 't Huys van Oostenryk, en de Gastanaga, is gisteren na Cadix t'zeyl gegaen.

Brussel den 18 July. Sondag trokken 2 Brandenburgse Regimenten te voet na 't Campement van den Marquis de Gastanaga, dat nu maer een uur van dese Stad leyt, en Maendag 't Regiment Dragonders van

Example ABBYY OCR Output



Uyt het Leger tot Gembloers den 16 Iuly. Heden deſe morgen arri- veerde weder een Courier, afgeſonden door den Heer Keurvorſt van Saxen, aen zyn Brittaniſche Majeſteyt, met tyding dat hy Heer Keur- vorſt de Franſſen in d'Arrier-Guarde was gevallen, en een Conſide- rabele avantagie hadde gekregen.

Aelſt den 17 Iuly. Syn Excell: onſe Gouverneur legt met ſijn Deta- chement nog tot Haren, tuſſchen Bruſſel en Vilvoorden. De Graef van Bergeyk is nu alleen tot Intendant Generael gemaekt, om 't Geld te ontfangen, en de Spaenſche Militie alle Maenden te betalen.

Gent den 18 Iuly. Op Sondag quam hier een Franſch Tamboer, om de party Franſſen, laetſt hier ingebragt, te rantſoeneeren; 't welk ge- ſchiedde. Doch la Fleur wierd giſteren hier opgehangen, voor wien ſy al 300 Piſtolen hadden geboden. Na dat de Gerantſoeneerden buy- ten de Poort waren, deſerteerden 'er met der haeſt tien.

Oſtende den 18 Iuly. Ons Convoy na Spanjen beſtaende uyt 17 Schepen, waer onder de Maegd van Gent, de St. Jacob, 't Huys van Ooſtenryk, en de Gaſtanaga, is giſteren na Cadix t'zeyl gegaen.

Bruſſel den 18 Iuly. Sondag trokken 2 Brandenburgſe Regimenten te voet na 't Campement van den Marquis de Gaſtanaga, dat nu maer een uur van deſe Stad leyt, en Maendag 't Regiment Dragonders van

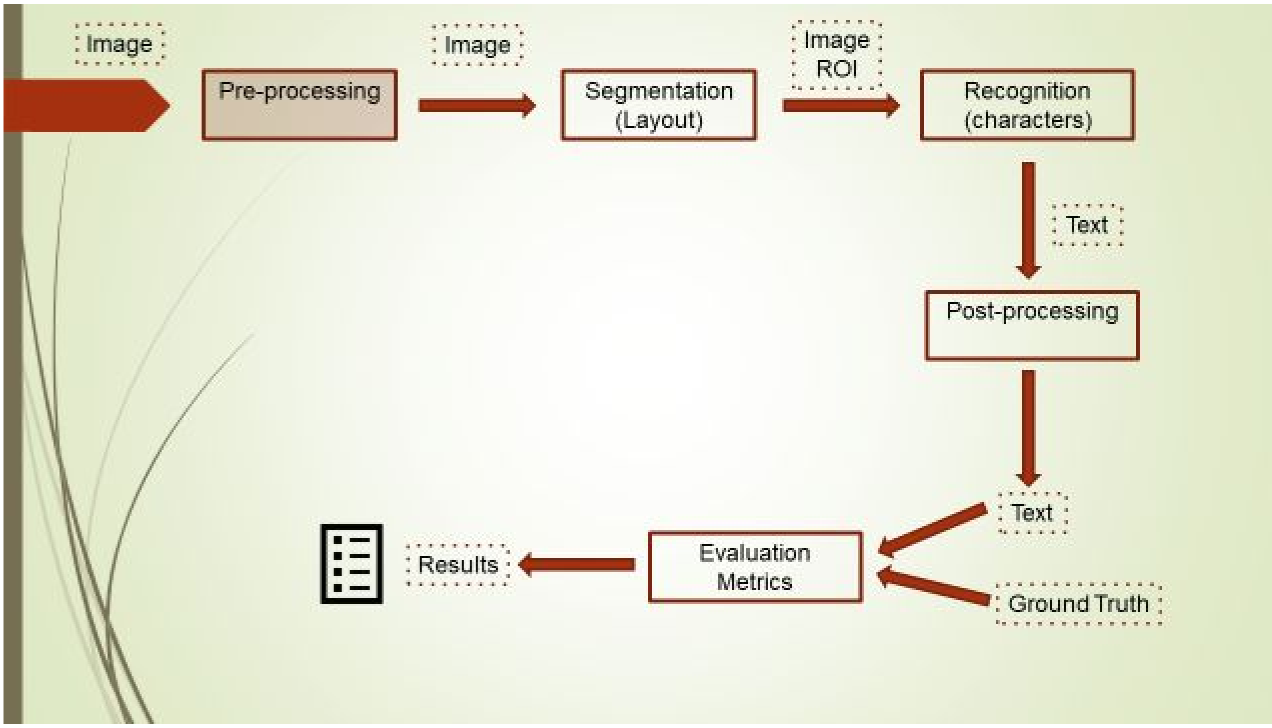
Vyt het Irg'r tot Gernhloers den 16 Inly. Heden defe morgen arrlâ weder een Courier, afgefonden door den Heer Kcutvotft van Saxen, acn zyn BrittanifÃ¨he Majefteyt, met tyding dat hy Heer KeÃ¼rvoitl de Franffcn in d'Arricr-Guarde was gevallen , en een Conlidcrabele avantagichadde gekregen. Ar!]} den 17 luly. Syn ExcÃ¨ll: onfe Gouverneur legt met fijn Detachement nog tot Haren , ttiffclisn Bruffel en Vilvoorden. De 6raef| van Bergeyk is nu alleen tot_Intendant Gcnerael gemaekt, om 't Geld te ontfangea, en de Spaeniche Militie alle Maenden te betalen. Geit den 1S luly. Op Sondag quain hier een Fianfch Tamboer, om de party Franffcn, betll hier ingebragt, te rantfoeneeren ; 't welk ge- ' fchiedde. Doch la Fleur wierd gifteren hier opgehangen, voor uien fy al 30- Piftolen hadden geboden, Na dat de Gerantf'oeneerden buytcn de Pooit waren, deferteerden 'et met der liaeft tien. Oofte.tdi Jen 18 Ittly. ()ns Convoy na Spanjen bellaende uyt 17 Schepen, waer onder de Maegd van Gent, de St. Jacob , 't Huys van OoÃ¶enryk, en de Gaftanaga, is gifteren na Cadix t'zeyl gegaen.

ABBYY OCR Output

Vyt het Irg'r tot Gernhloers den 16 Inly. Heden defe morgen arrlâ weder een Courier, afgefonden door den Heer Kcutvotft van Saxen, acn zyn BrittanifÃ¨he Majefteyt, met tyding dat hy Heer KeÃ¼rvoitl de Franffcn in d'Arricr-Guarde was gevallen , en een Conlidcrabele avantagichadde gekregen. Ar!]} den 17 luly. Syn ExcÃ¨ll: onfe Gouverneur legt met fijn Detachement nog tot Haren , ttiffclisn Bruffel en Vilvoorden. De 6raef| van Bergeyk is nu alleen tot_Intendant Gcnerael gemaekt, om 't Geld te ontfangea, en de Spaeniche Militie alle Maenden te betalen. Geit den 1S luly. Op Sondag quain hier een Fianfch Tamboer, om de party Franffcn, betll hier ingebragt, te rantfoeneeren ; 't welk ge- ' fchiedde. Doch la Fleur wierd gifteren hier opgehangen, voor uien fy al 30- Piftolen hadden geboden, Na dat de Gerantf'oeneerden buytcn de Pooit waren, deferteerden 'et met der liaeft tien. Oofte.tdi Jen 18 Ittly. ()ns Convoy na Spanjen bellaende uyt 17 Schepen, waer onder de Maegd van Gent, de St. Jacob , 't Huys van OoÃ¶enryk, en de Gaftanaga, is gifteren na Cadix t'zeyl gegaen.

Ground Truth

Uyt het Leger tot Gembloers den 16 luly. Heden dese morgen arri veerde weder een Courier, afgesonden door den Heer Keurvorst van Saxen, aen zyn Brittanische Majesteyt, met tyding dat hy Heer Keurvorst de Franssen in d'Arrier-Guarde was gevallen, en een Considerabele avantagie hadde gekregen. Aalst den 17 luly. Syn Excell: onse Gouverneur legt met sijn Detachement nog tot Haren, tusschen Brussel en Vilvoorden. De Graef van Bergeyk is nu alleen tot Intendant Generael gemaekt, om 't Geld te ontfangen, en de Spaensche Militie alle Maenden te betalen. Gent den 18 luly. Op Sondag quam hier een Fransch Tamboer, om de party Franssen, laetst hier ingebragt, te rantsoeneeren; 't welk geschiedde. Doch la Fleur wierd gisteren hier opgehangen, voor wien sy al 300 Pistolen hadden geboden, Na dat de Gerantsoeneerden buyten de Poort waren, deserteerden 'er met der haest tien. Oostende den 18 luly. Ons Convoy na Spanjen bestaende uyt 17 Schepen, waer onder de Maegd van Gent, de St. Jacob, 't Huys van Oostenryk, en de Gastanaga, is gisteren na Cadix t'zeyl gegaen.



Preprocessing & Layout

Xue Wang
Jerry Guo

Xue Wang

together with Rutger & Lambert

Bottom-up word segmentation using a pixel to pixel deep-learning tool

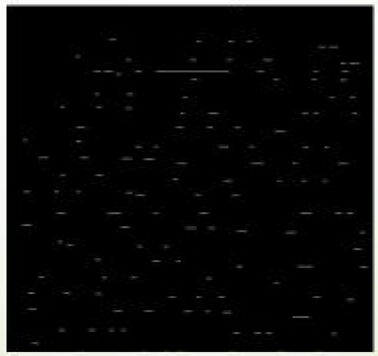
- Goal: Model --->find the word position & spaces in the image
- Dataset: image & alto.xml (with annotation about spaces and text strings)
- Step 1: clean the data and make the ground truth

Raw image

VRANCKRYCK.

PAris, den 12 Augusti. Den Hertog van Sunderlant continueert hier in besoignes, doende groote klachten over onse Fregatten, die hebben geweygert te strijcken voor sijn Kon. Vlagge, maer vint weynig gehoor, als ook over de sollicitatie voor den Hertog van Lotteringen, want sijn Majesteyt oordeelende van dat Huys te veel schade en affront geleden te hebben, wil buyten de presentatie niets doen in sijn voordeel. Den Heere Gravenbal doet insgelijcks groote devoir tot de evacuatie van de Steden, waer in sijne Majesteyt ook toegestaen heeft; doch onder voorwaerden van 2 van de voornaemste Plaetsen in sijn eygen bewaring te houden, tot'er tijt dat Brandenburg behoorlijcke satisfactie aen Sweden sal gegeven hebben. Bergen is naeu beset, en alle Advenues soodanig gefortificeert, dat men niet vertrouwt, dat de Geallieerden yets sullen onderstaen om daer in te brengen. De schade geleden van de 7 Oorlogschepen en 5 Fluyten, onder den Grave d'Estre, by 't Eylant Aves, wort begroot op 6 Millioenen, en wert de directie van dien Grave niet wel opgenomen.

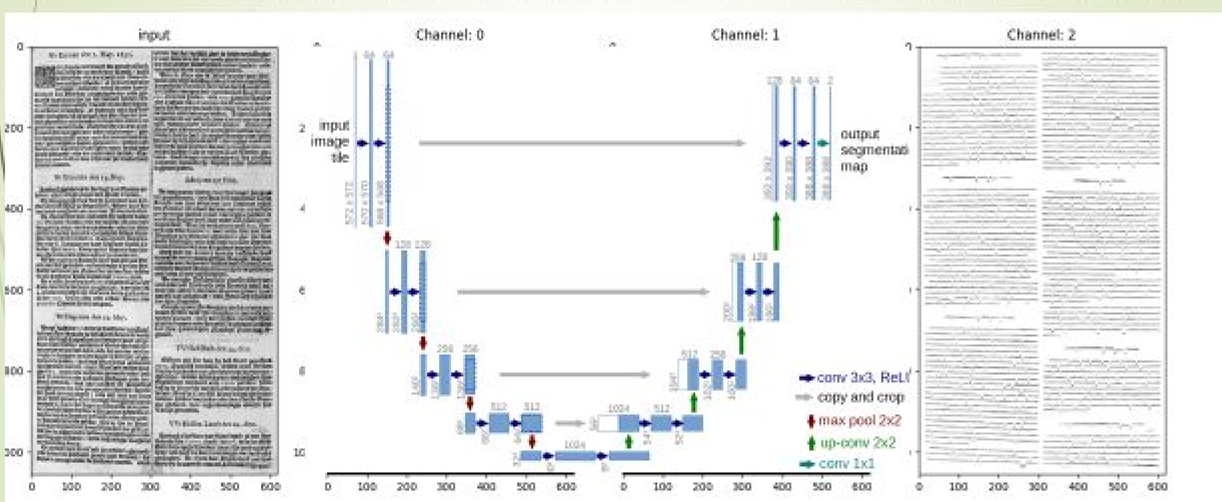
Ground-truth:space



Ground-truth:string

Xue Wang

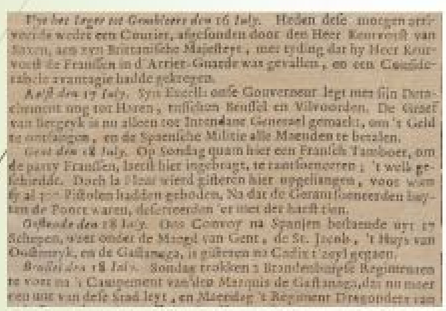
- Train Model: based on ARU-Net
  (output: the word and spaces in the image!)
  ➔ sometimes better than ABBY because directly based on pixels



Preprocessing

- Region information:
  - we took a shortcut...
- We apply different binarization methods
  - Local
  - Global
  - Both on plain grayscale input and DNN enhanced documents
- We don't have ground truth for binarization
- We decide to compare the OCR results directly
  - Tesseract

Preprocessing



Uyt het Leger tot Gembloers den 16 July. Heden dese morgen arriveerde weder een Courier, afgesonden door den Heer Keurvorst van Saxen, aen syn Britannische Majesteyt, met tyding dat hy Heer Keurvorst de Fransen in d'Arrier-Guarde was gevallen, en een Considerabele avantagie hadde gekregen.

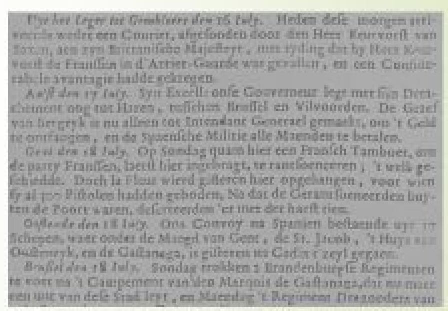
Aalst den 17 July. Syn Excell: onse Gouverneur legt met sijn Detachement nog tot Haren, tusschen Brussel en Vilvoorden. De Graef van Bergeyk is nu alleen tot Intendant Generael gemaekt, om 't Geld te ontfangen, en de Spaensche Militie alle Maenden te betalen.

Gent den 18 July. Op Sondag quam hier een Fransch Tamboer, om de party Fransen, laetst hier ingebragt, te rantsoeneeren; 't welk geschiedde. Doch la Fleur wierd gisteren hier opgehangen, voor wien sy al 500 Pistolen hadden geboden, Na dat de Garantsoeneerden buyten de Poort waren, deserteerden 'er met der haest tien.

Ostende den 18 July. Ons Convoy na Spanjen bestaende uyt 17 Schepen, waer onder de Maegd van Gent, de St. Jacob, 't Huys van Oostenryk, en de Gastanaga, is gisteren na Cadix t'zeyl gegaen.

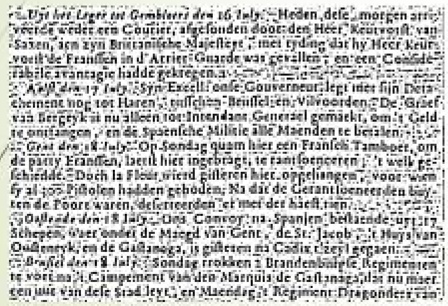
Brussel den 18 July. Sondag trokken 2 Brandenburgse Regimenten te voet na 't Campement van den Marquis de Gastanaga, dat nu meer een uur van dese Stad leyt, en Maendag 't Regiment Dragonders van

Plain input



Enhanced by *DeepOtsu*

Preprocessing



Uyt het Leger tot Gemblours den 16 July. Heden dese morgen arriveerde weder een Courier, afgesonden door den Heer Keurvorst van Saxen, aen zyn Brittanische Majesteyt, met tyding dat hy Heer Keurvorst de Franssen in d'Arrier-Guarde was gevallen, en een Considerabele avantagie hadde gekregen.

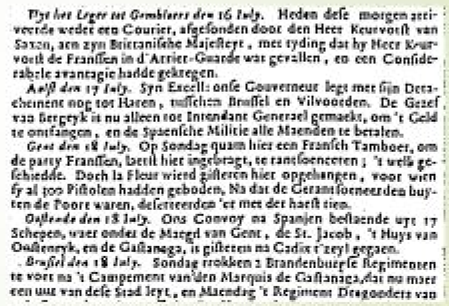
Aelst den 17 July. Syn Excell: onse Gouverneur legt met sijn Detachement nog tot Haren, tusschen Brussel en Vilvoorden. De Graef van Bergeyk is nu alleen tot Intendant Generael gemaekt, om 't Geld te ontfangen, en de Spaensche Militie alle Maenden te betalen.

Gent den 18 July. Op Sondag quam hier een Fransch Tamboer, om de party Franssen, laetst hier ingebragt, te rantsoeneeren; 't welk geschiedde. Doch la Fleur wierd gisteren hier opgehangen, voor wien sy al 300 Pistolen hadden geboden, Na dat de Gerantsoeneerden buyten de Poort waren, deserteerden 'er met der haest tien.

Ostende den 18 July. Ons Convoy na Spanjen bestaende uyt 17 Schepen, waer onder de Maegd van Gent, de St. Jacob, 't Huys van Oostenryk, en de Gastanaga, is gisteren na Cadix t'zeyl gegaen.

Brussel den 18 July. Sondag trokken 2 Brandenburgse Regimenten te voet na 't Campement van den Marquis de Gastanaga, dat nu maer een uur van dese Stad leyt, en Maendag 't Regiment Dragonders van

Plain input
+ adaptive mean



Uyt het Leger tot Gemblours den 16 July. Heden dese morgen arriveerde weder een Courier, afgesonden door den Heer Keurvorst van Saxen, aen zyn Brittanische Majesteyt, met tyding dat hy Heer Keurvorst de Franssen in d'Arrier-Guarde was gevallen, en een Considerabele avantagie hadde gekregen.

Aelst den 17 July. Syn Excell: onse Gouverneur legt met sijn Detachement nog tot Haren, tusschen Brussel en Vilvoorden. De Graef van Bergeyk is nu alleen tot Intendant Generael gemaekt, om 't Geld te ontfangen, en de Spaensche Militie alle Maenden te betalen.

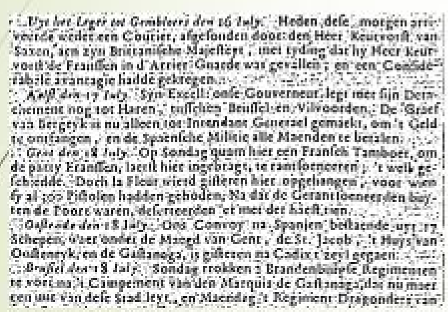
Gent den 18 July. Op Sondag quam hier een Fransch Tamboer, om de party Franssen, laetst hier ingebragt, te rantsoeneeren; 't welk geschiedde. Doch la Fleur wierd gisteren hier opgehangen, voor wien sy al 300 Pistolen hadden geboden, Na dat de Gerantsoeneerden buyten de Poort waren, deserteerden 'er met der haest tien.

Ostende den 18 July. Ons Convoy na Spanjen bestaende uyt 17 Schepen, waer onder de Maegd van Gent, de St. Jacob, 't Huys van Oostenryk, en de Gastanaga, is gisteren na Cadix t'zeyl gegaen.

Brussel den 18 July. Sondag trokken 2 Brandenburgse Regimenten te voet na 't Campement van den Marquis de Gastanaga, dat nu maer een uur van dese Stad leyt, en Maendag 't Regiment Dragonders van

Enhanced by *DeepOtsu*
+ adaptive mean

Preprocessing



Uyt het Leger tot Gemblours den 16 July. Heden dese morgen arriveerde weder een Courier, afgesonden door den Heer Keurvorst van Saxen, aen zyn Brittanische Majesteyt, met tyding dat hy Heer Keurvorst de Franssen in d'Arrier-Guarde was gevallen, en een Considerabele avantagie hadde gekregen.

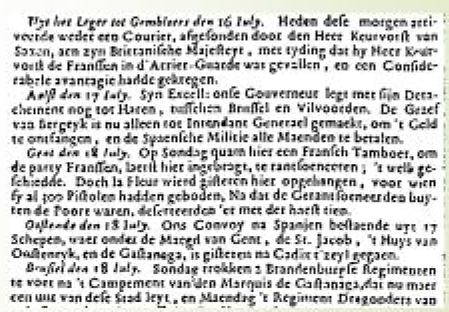
Aelst den 17 July. Syn Excell: onse Gouverneur legt met sijn Detachement nog tot Haren, tusschen Brussel en Vilvoorden. De Graef van Bergeyk is nu alleen tot Intendant Generael gemaekt, om 't Geld te ontfangen, en de Spaensche Militie alle Maenden te betalen.

Gent den 18 July. Op Sondag quam hier een Fransch Tamboer, om de party Franssen, laetst hier ingebragt, te rantsoeneeren; 't welk geschiedde. Doch la Fleur wierd gisteren hier opgehangen, voor wien sy al 300 Pistolen hadden geboden, Na dat de Gerantsoeneerden buyten de Poort waren, deserteerden 'er met der haest tien.

Ostende den 18 July. Ons Convoy na Spanjen bestaende uyt 17 Schepen, waer onder de Maegd van Gent, de St. Jacob, 't Huys van Oostenryk, en de Gastanaga, is gisteren na Cadix t'zeyl gegaen.

Brussel den 18 July. Sondag trokken 2 Brandenburgse Regimenten te voet na 't Campement van den Marquis de Gastanaga, dat nu maer een uur van dese Stad leyt, en Maendag 't Regiment Dragonders van

Plain input
+ adaptive gaussian



Uyt het Leger tot Gemblours den 16 July. Heden dese morgen arriveerde weder een Courier, afgesonden door den Heer Keurvorst van Saxen, aen zyn Brittanische Majesteyt, met tyding dat hy Heer Keurvorst de Franssen in d'Arrier-Guarde was gevallen, en een Considerabele avantagie hadde gekregen.

Aelst den 17 July. Syn Excell: onse Gouverneur legt met sijn Detachement nog tot Haren, tusschen Brussel en Vilvoorden. De Graef van Bergeyk is nu alleen tot Intendant Generael gemaekt, om 't Geld te ontfangen, en de Spaensche Militie alle Maenden te betalen.

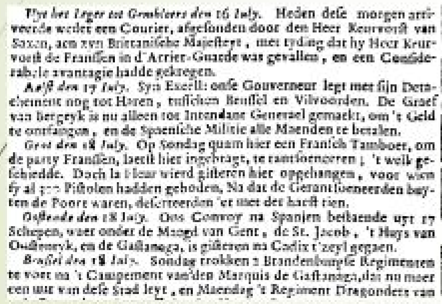
Gent den 18 July. Op Sondag quam hier een Fransch Tamboer, om de party Franssen, laetst hier ingebragt, te rantsoeneeren; 't welk geschiedde. Doch la Fleur wierd gisteren hier opgehangen, voor wien sy al 300 Pistolen hadden geboden, Na dat de Gerantsoeneerden buyten de Poort waren, deserteerden 'er met der haest tien.

Ostende den 18 July. Ons Convoy na Spanjen bestaende uyt 17 Schepen, waer onder de Maegd van Gent, de St. Jacob, 't Huys van Oostenryk, en de Gastanaga, is gisteren na Cadix t'zeyl gegaen.

Brussel den 18 July. Sondag trokken 2 Brandenburgse Regimenten te voet na 't Campement van den Marquis de Gastanaga, dat nu maer een uur van dese Stad leyt, en Maendag 't Regiment Dragonders van

Enhanced by *DeepOtsu*
+ adaptive gaussian

Preprocessing



Uyt het Leger tot Gemblours den 16 July. Heden dese morgen arriveerde wedet een Courier, afgesonden door den Heer Keurvorst van Saxen, aen zyn Brittanische Majesteyt, met tyding dat hy Heer Keurvorst de Franssen in d'Arrier-Guarde was gevallen, en een Considerable avantagie hadde gekregen.

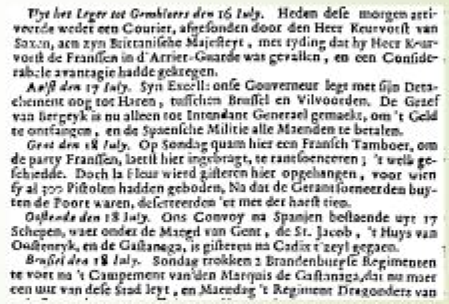
Aelst den 17 July. Syn Excell: onse Gouverneur legt met sijn Detachement nog tot Haren, tusschen Brussel en Vilvoorden. De Graef van Bergeyk is nu alleen tot Intendant Generael gemaekt, om 't Geld te ontfangen, en de Spaensche Militie alle Maenden te betalen.

Gent den 18 July. Op Sondag quam hier een Fransch Tamboer, om de party Franssen, laetst hier ingebragt, te ranssoeneeren; 't welk geschiedde. Doch la Fleur wierd gisteren hier opgehangen, voor wien sy al 300 Pistolen hadden geboden, Na dat de Gerantsoeneerden buyten de Poort waren, deserteerden 'er met der haest tien.

Ostende den 18 July. Ons Convoy na Spanjen bestaende uyt 17 Schepen, waer onder de Maegd van Gent, de St. Jacob, 't Huys van Oostenryk, en de Gastanaga, is gisteren na Cadix t'zeyl gegaen.

Brussel den 18 July. Sondag trokken 2 Brandenburgse Regimenten te voet na 't Campement van den Marquis de Gastanaga, dat nu meer een uur van dese Stad leyt, en Maendag 't Regiment Dragonders van

Plain input
+ Niblack



Uyt het Leger tot Gemblours den 16 July. Heden dese morgen arriveerde wedet een Courier, afgesonden door den Heer Keurvorst van Saxen, aen zyn Brittanische Majesteyt, met tyding dat hy Heer Keurvorst de Franssen in d'Arrier-Guarde was gevallen, en een Considerable avantagie hadde gekregen.

Aelst den 17 July. Syn Excell: onse Gouverneur legt met sijn Detachement nog tot Haren, tusschen Brussel en Vilvoorden. De Graef van Bergeyk is nu alleen tot Intendant Generael gemaekt, om 't Geld te ontfangen, en de Spaensche Militie alle Maenden te betalen.

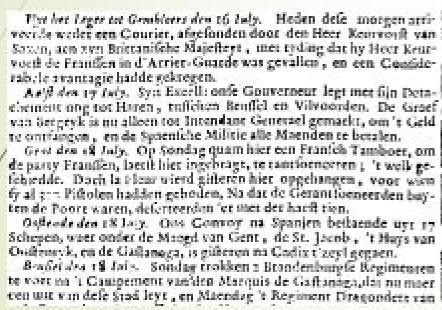
Gent den 18 July. Op Sondag quam hier een Fransch Tamboer, om de party Franssen, laetst hier ingebragt, te ranssoeneeren; 't welk geschiedde. Doch la Fleur wierd gisteren hier opgehangen, voor wien sy al 300 Pistolen hadden geboden, Na dat de Gerantsoeneerden buyten de Poort waren, deserteerden 'er met der haest tien.

Ostende den 18 July. Ons Convoy na Spanjen bestaende uyt 17 Schepen, waer onder de Maegd van Gent, de St. Jacob, 't Huys van Oostenryk, en de Gastanaga, is gisteren na Cadix t'zeyl gegaen.

Brussel den 18 July. Sondag trokken 2 Brandenburgse Regimenten te voet na 't Campement van den Marquis de Gastanaga, dat nu meer een uur van dese Stad leyt, en Maendag 't Regiment Dragonders van

Enhanced by *DeepOtsu*
+ Niblack

Preprocessing



Uyt het Leger tot Gemblours den 16 July. Heden dese morgen arriveerde wedet een Courier, afgesonden door den Heer Keurvorst van Saxen, aen zyn Brittanische Majesteyt, met tyding dat hy Heer Keurvorst de Franssen in d'Arrier-Guarde was gevallen, en een Considerable avantagie hadde gekregen.

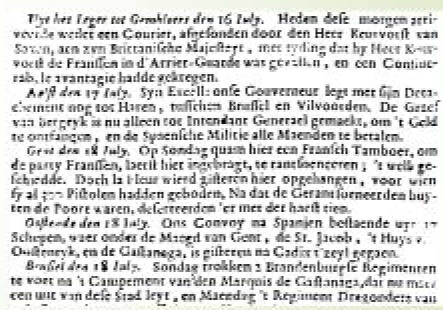
Aelst den 17 July. Syn Excell: onse Gouverneur legt met sijn Detachement nog tot Haren, tusschen Brussel en Vilvoorden. De Graef van Bergeyk is nu alleen tot Intendant Generael gemaekt, om 't Geld te ontfangen, en de Spaensche Militie alle Maenden te betalen.

Gent den 18 July. Op Sondag quam hier een Fransch Tamboer, om de party Franssen, laetst hier ingebragt, te ranssoeneeren; 't welk geschiedde. Doch la Fleur wierd gisteren hier opgehangen, voor wien sy al 300 Pistolen hadden geboden, Na dat de Gerantsoeneerden buyten de Poort waren, deserteerden 'er met der haest tien.

Ostende den 18 July. Ons Convoy na Spanjen bestaende uyt 17 Schepen, waer onder de Maegd van Gent, de St. Jacob, 't Huys van Oostenryk, en de Gastanaga, is gisteren na Cadix t'zeyl gegaen.

Brussel den 18 July. Sondag trokken 2 Brandenburgse Regimenten te voet na 't Campement van den Marquis de Gastanaga, dat nu meer een uur van dese Stad leyt, en Maendag 't Regiment Dragonders van

Plain input
+ Sauvola



Uyt het Leger tot Gemblours den 16 July. Heden dese morgen arriveerde wedet een Courier, afgesonden door den Heer Keurvorst van Saxen, aen zyn Brittanische Majesteyt, met tyding dat hy Heer Keurvorst de Franssen in d'Arrier-Guarde was gevallen, en een Considerable avantagie hadde gekregen.

Aelst den 17 July. Syn Excell: onse Gouverneur legt met sijn Detachement nog tot Haren, tusschen Brussel en Vilvoorden. De Graef van Bergeyk is nu alleen tot Intendant Generael gemaekt, om 't Geld te ontfangen, en de Spaensche Militie alle Maenden te betalen.

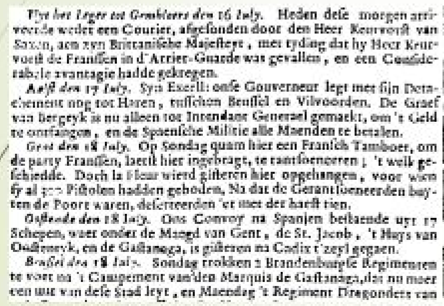
Gent den 18 July. Op Sondag quam hier een Fransch Tamboer, om de party Franssen, laetst hier ingebragt, te ranssoeneeren; 't welk geschiedde. Doch la Fleur wierd gisteren hier opgehangen, voor wien sy al 300 Pistolen hadden geboden, Na dat de Gerantsoeneerden buyten de Poort waren, deserteerden 'er met der haest tien.

Ostende den 18 July. Ons Convoy na Spanjen bestaende uyt 17 Schepen, waer onder de Maegd van Gent, de St. Jacob, 't Huys van Oostenryk, en de Gastanaga, is gisteren na Cadix t'zeyl gegaen.

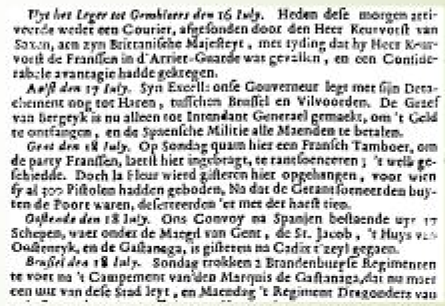
Brussel den 18 July. Sondag trokken 2 Brandenburgse Regimenten te voet na 't Campement van den Marquis de Gastanaga, dat nu meer een uur van dese Stad leyt, en Maendag 't Regiment Dragonders van

Enhanced by *DeepOtsu*
+ Sauvola

Preprocessing

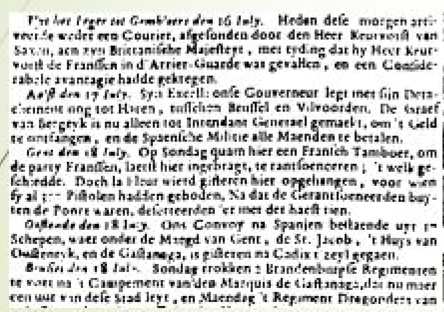

Uyt het Leger tot Gembloers den 16 July. Heden dese morgen arriveerde weder een Courier, afgesonden door den Heer Keurvorst van Saxen, aen zyn Brittanische Majesteyt, met tyding dat hy Heer Keurvorst de Franssen in d'Arrier-Guarde was gevallen, en een Considerabele avantagie hadde gekregen.
Aelst den 17 July. Syn Excell: onse Gouverneur legt met sijn Detachement nog tot Haren, tusschen Brussel en Vilvoorden. De Graef van Bergeyk is nu alleen tot Intendant Generael gemaekt, om 't Geld te ontfangen, en de Spaensche Militie alle Maenden te betalen.
Gent den 18 July. Op Sondag quam hier een Fransch Tamboer, om de party Franssen, laetst hier ingebragt, te rantsoeneeren; 't welk geschiedde. Doch la Fleur wierd gisteren hier opgehangen, voor wien sy al 300 Pistolen hadden geboden, Na dat de Gerantsoeneerden buyten de Poort waren, deserteerden 'er met der haest tien.
Ostende den 18 July. Ons Convoy na Spanjen bestaende uyt 17 Schepen, waer onder de Maegd van Gent, de St. Jacob, 't Huys van Oostenryk, en de Gastanega, is gisteren na Cadix t'zeyl gegaen.
Brussel den 18 July. Sondag trokken 2 Brandenburgse Regimenten te voet na 't Campement van den Marquis de Gastanaga, dat nu meer een uur van dese Stad leyt, en Maendag 't Regiment Dragonders van

Plain input
+ Otsu

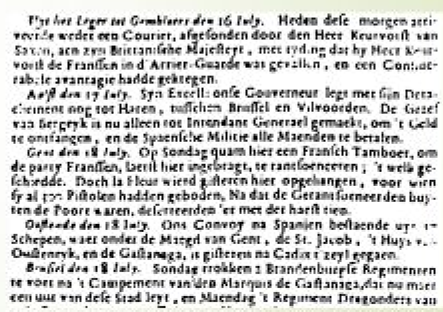

Uyt het Leger tot Gembloers den 16 July. Heden dese morgen arriveerde weder een Courier, afgesonden door den Heer Keurvorst van Saxen, aen zyn Brittanische Majesteyt, met tyding dat hy Heer Keurvorst de Franssen in d'Arrier-Guarde was gevallen, en een Considerabele avantagie hadde gekregen.
Aelst den 17 July. Syn Excell: onse Gouverneur legt met sijn Detachement nog tot Haren, tusschen Brussel en Vilvoorden. De Graef van Bergeyk is nu alleen tot Intendant Generael gemaekt, om 't Geld te ontfangen, en de Spaensche Militie alle Maenden te betalen.
Gent den 18 July. Op Sondag quam hier een Fransch Tamboer, om de party Franssen, laetst hier ingebragt, te rantsoeneeren; 't welk geschiedde. Doch la Fleur wierd gisteren hier opgehangen, voor wien sy al 300 Pistolen hadden geboden, Na dat de Gerantsoeneerden buyten de Poort waren, deserteerden 'er met der haest tien.
Ostende den 18 July. Ons Convoy na Spanjen bestaende uyt 17 Schepen, waer onder de Maegd van Gent, de St. Jacob, 't Huys van Oostenryk, en de Gastanega, is gisteren na Cadix t'zeyl gegaen.
Brussel den 18 July. Sondag trokken 2 Brandenburgse Regimenten te voet na 't Campement van den Marquis de Gastanaga, dat nu meer een uur van dese Stad leyt, en Maendag 't Regiment Dragonders van

Enhanced by *DeepOtsu*
+ Otsu

Preprocessing


Uyt het Leger tot Gembloers den 16 July. Heden dese morgen arriveerde weder een Courier, afgesonden door den Heer Keurvorst van Saxen, aen zyn Brittanische Majesteyt, met tyding dat hy Heer Keurvorst de Franssen in d'Arrier-Guarde was gevallen, en een Considerabele avantagie hadde gekregen.
Aelst den 17 July. Syn Excell: onse Gouverneur legt met sijn Detachement nog tot Haren, tusschen Brussel en Vilvoorden. De Graef van Bergeyk is nu alleen tot Intendant Generael gemaekt, om 't Geld te ontfangen, en de Spaensche Militie alle Maenden te betalen.
Gent den 18 July. Op Sondag quam hier een Fransch Tamboer, om de party Franssen, laetst hier ingebragt, te rantsoeneeren; 't welk geschiedde. Doch la Fleur wierd gisteren hier opgehangen, voor wien sy al 300 Pistolen hadden geboden, Na dat de Gerantsoeneerden buyten de Poort waren, deserteerden 'er met der haest tien.
Ostende den 18 July. Ons Convoy na Spanjen bestaende uyt 17 Schepen, waer onder de Maegd van Gent, de St. Jacob, 't Huys van Oostenryk, en de Gastanega, is gisteren na Cadix t'zeyl gegaen.
Brussel den 18 July. Sondag trokken 2 Brandenburgse Regimenten te voet na 't Campement van den Marquis de Gastanaga, dat nu meer een uur van dese Stad leyt, en Maendag 't Regiment Dragonders van

Plain input
+ Otsu blurred input


Uyt het Leger tot Gembloers den 16 July. Heden dese morgen arriveerde weder een Courier, afgesonden door den Heer Keurvorst van Saxen, aen zyn Brittanische Majesteyt, met tyding dat hy Heer Keurvorst de Franssen in d'Arrier-Guarde was gevallen, en een Considerabele avantagie hadde gekregen.
Aelst den 17 July. Syn Excell: onse Gouverneur legt met sijn Detachement nog tot Haren, tusschen Brussel en Vilvoorden. De Graef van Bergeyk is nu alleen tot Intendant Generael gemaekt, om 't Geld te ontfangen, en de Spaensche Militie alle Maenden te betalen.
Gent den 18 July. Op Sondag quam hier een Fransch Tamboer, om de party Franssen, laetst hier ingebragt, te rantsoeneeren; 't welk geschiedde. Doch la Fleur wierd gisteren hier opgehangen, voor wien sy al 300 Pistolen hadden geboden, Na dat de Gerantsoeneerden buyten de Poort waren, deserteerden 'er met der haest tien.
Ostende den 18 July. Ons Convoy na Spanjen bestaende uyt 17 Schepen, waer onder de Maegd van Gent, de St. Jacob, 't Huys van Oostenryk, en de Gastanega, is gisteren na Cadix t'zeyl gegaen.
Brussel den 18 July. Sondag trokken 2 Brandenburgse Regimenten te voet na 't Campement van den Marquis de Gastanaga, dat nu meer een uur van dese Stad leyt, en Maendag 't Regiment Dragonders van

Enhanced by *DeepOtsu*
+ Otsu blurred input

Preprocessing

OCR: Ground truth

Uyt het Leger tot Gembloers den 16 luly. Heden dese morgen arri
veerde weder een Courier, afgesonden door den Heer Keurvors
t van Saxen, aen zyn Brittanische Majesteyt, met tyding dat hy H
eer Keurvorst de Franssen in d'Arrier-
Guarde was gevallen, en een Considerabele avantagie hadde
gekregen. Aalst den 17 luly. Syn Excell: onse Gouverneur legt me
t sijn Detachement nog tot Haren, tusschen Brussel en Vilvoorden
. De Graef van Bergeyk is nu alleen tot Intendant Generael gem
aekt, om 't Geld te ontfangen, en de Spaensche Militie alle Mae
nden te betalen. …

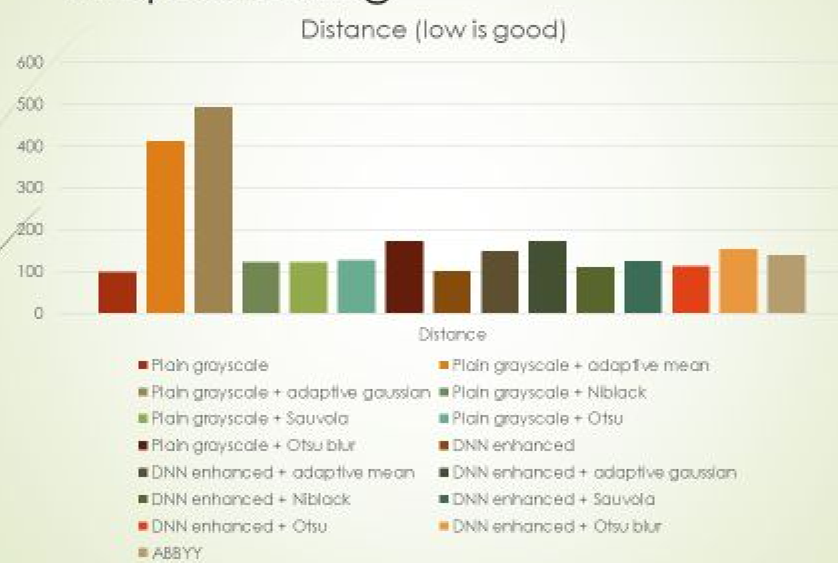
OCR: ABBYY

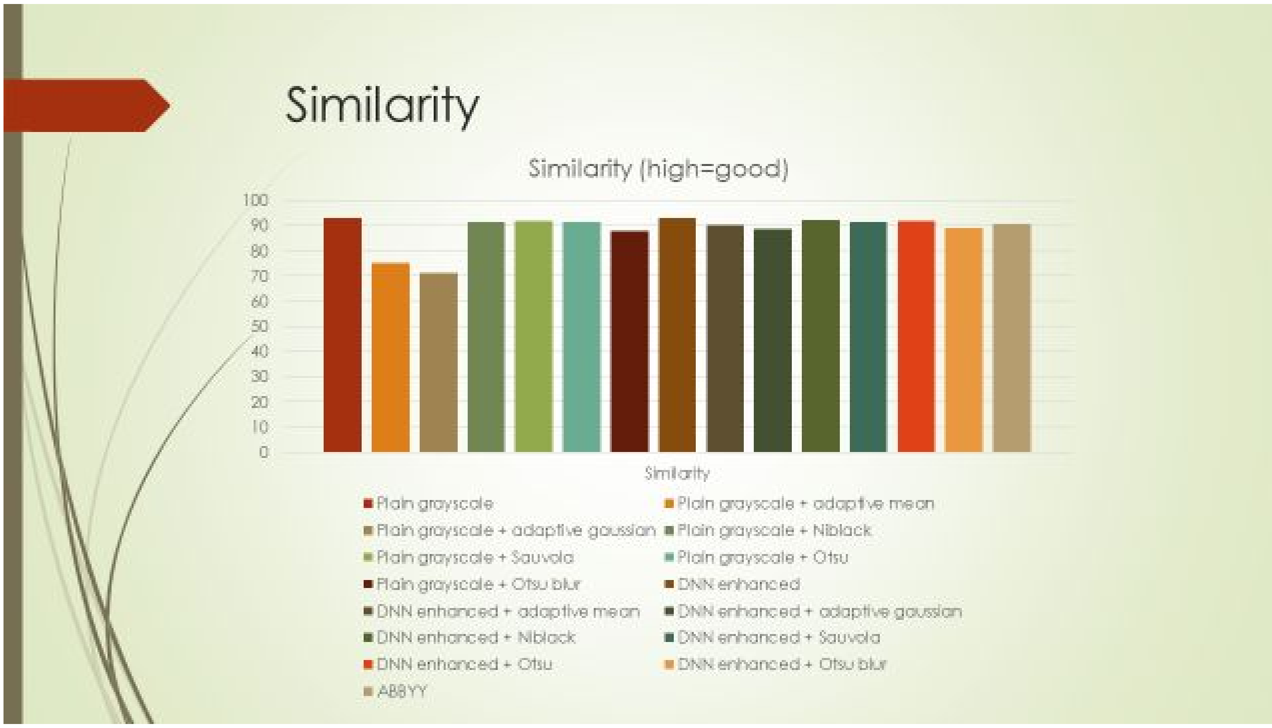
Vyt het lrg'r tot Gemhloers den 16 lnly. Heden defe morgen arriâ\x
96\xa0 weder een Courier, afgefonden door den Heer Kcutvotft v
an Saxen, acn zyn BrittanitÃ¯he Majefteyt, met tyding dat hy Heer
KeÃ¼rvoltl de Franffcn in d'Arricr-
Guarde was gevallen , en een Confidcrabele avantagichadde g
ekregen. Ar]] den 17 luly. Syn ExcÃ¨ll: onte Gouverneur legt met fij
n Detachement nog tot Haren , Hiffclsn Bruffel en Vilvoorden. De 6
raef | van Bergeyk is nu alleen tot_Intendant Gcneraal gemaekt, o
m 't Geld te ontfangea, en de Spaeniche Militie alle Maenden te
betalen. …

Plain + Otsu

Tlyt bet Leger tot Gembloers den 16
twly. Heden defe morgen arti-
| veerde weder een Courier, afgef
onden door den Heer Keurvorft van
Saxen, aen zyn Brittanifehe Majefte
ye , met tyding dat hy Heer Keus-
vorit de Franflen in d' Arrier-
Guarde was gevallen, en een Cont
Ide-
. rabzle avantagle hadde gekrege
n. : Aalf den 07 lnly. Syn Excell: onfe
Gouverneur legt …

Preprocessing




Similarity
Similarity (high=good)
100
90
80
70
60
50
40
30
20
10
0
Similarity
Plain grayscale
Plain grayscale + adaptive mean
Plain grayscale + adaptive gaussian
Plain grayscale + Niblack
Plain grayscale + Sauvola
Plain grayscale + Otsu
Plain grayscale + Otsu blur
DNN enhanced
DNN enhanced + adaptive mean
DNN enhanced + adaptive gaussian
DNN enhanced + Niblack
DNN enhanced + Sauvola
DNN enhanced + Otsu
DNN enhanced + Otsu blur
ABBYY

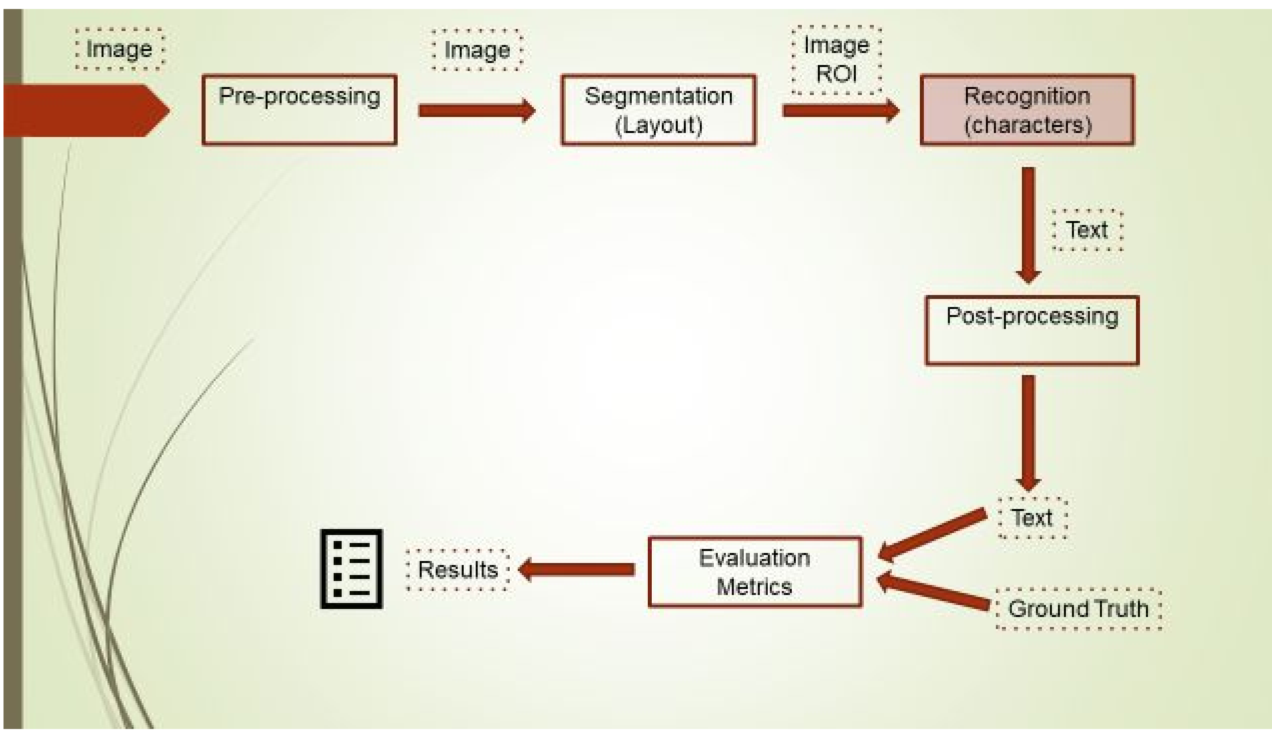
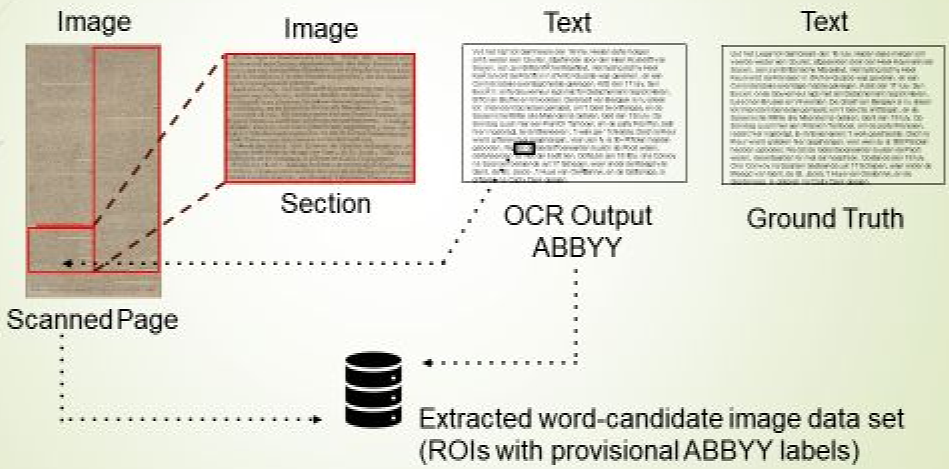


Image
Pre-processing
Image
Segmentation
(Layout)
Image
ROI
Recognition
(characters)
Text
Post-processing
Text
Ground Truth
Evaluation
Metrics
Results

Lambert

Data: Newspapers (1662 – 1795)



Mahya Ameryan
Lambert Schomaker

Text recognition

- **a) Tabula rasa start from zero labels on a Gothic-script collection: Monk**

vs

- **b) How to exploit the sub-optimal recognition results from commercial providers by using the image ROIs from the XML metadata**

Lambert Schomaker

Text recognition/a – Monk demo

- **Tabula rasa start from zero labels, on a Gothic-script collection: Monk**

- **436 scanned pages (Geleghentheyt 'sHertogen-Bosch)**
- **XML of commercial provider contains a lot of exotic unicode**
- **Normalized (flattened) to ASCII to facilitate performance evaluation**

- **Two labelers, not more than 2 hours of work**
- **Harvested 319 word classes, 1760 human-labeled images**
- **5.5M candidate images processed, learning is going on**

- **Method: Alternation between a recognition & retrieval engine: 'Fahrkunst' principle "HPC + Human in the loop"**

van Oosten & Schomaker (2014). Separability versus prototypicality in handwritten word-image retrieval, Pattern Recognition, 47(3), pp. 1031-1038

Lambert Schomaker

Monk

Mahya Ameryan
Lambert Schomaker

Text recognition/b: exploiting results of 'commercial' recognition attempts

Methods of labeling:

- Human labeling: high quality, but expensive.
- **Machine labeling: Maybe low quality, but massive data**

How to exploit the sub-optimal recognition results from commercial providers by using the image ROIs from the XML metadata?

- **Using the 'Meertens' data set:**
  scans + sub-optimal recognition results from the Alto XML format
- **Collect image ROIs with word candidates+ their provisional labels**

Data & result of Word recognition

Mahya Ameryan
Lambert Schomaker

	#Images	#Classes case sensitive	#Classes case insensitive
Train	10,037	701	654
Validation	3,346	665	618
Test	3,346	665	618
Total Dataset	16,728	736	689

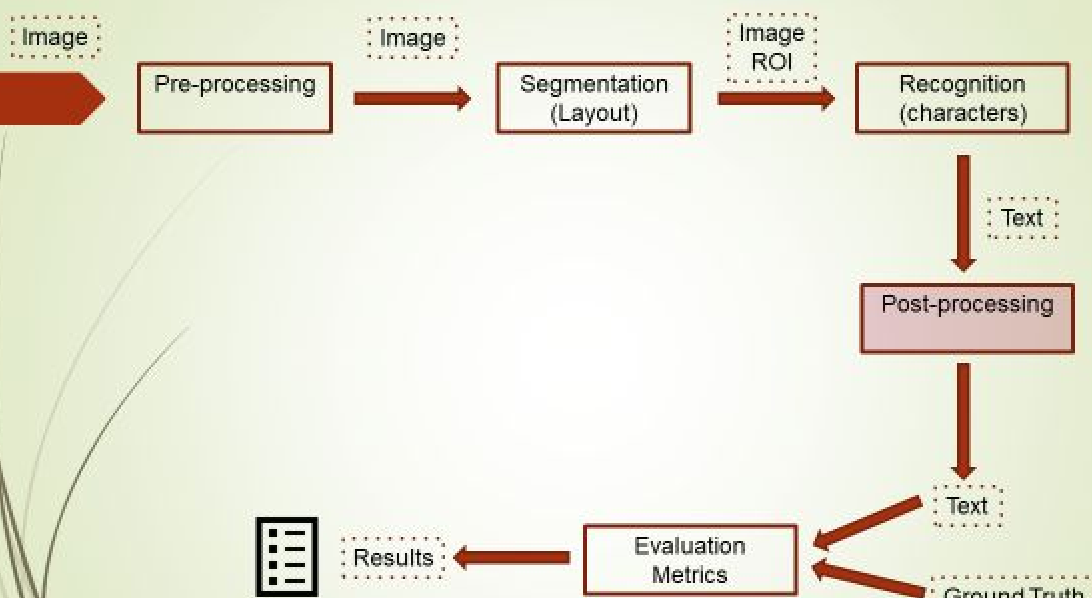
- **Alphabet size is 66; it includes [A-Z] [a-z], digits, and ^ ? = - . @**
-
- **A CNN/LSTM model was trained**
-
- **88% Word-recognition accuracy** **(while 100% can never be reached!)**

Mahya Ameryan & Lambert Schomaker (2019). A limited-size ensemble of homogeneous CNN/LSTMs for high-performance word classification, arXiv:1912.03223

Text recognition/conclusion

Mahya Ameryan
Lambert Schomaker

- In current practice, the ROIs (images) of recognized words are never used
  Our results show that this information is highly valuable, even with errors.
- Next steps:
  - Cleaning up the data: Filter out strange OCR results (words start with letters) and assume that a word string that happens >n times is not a coincidence: it *may* be a lexical word (n=5 or 10).
  - Applying the method on a more realistic 4894 (5K) classes
- Raw ROI-image extraction from .JPG/.XML is ongoing:
  22k classes, 78k instances on 24/1/2020, 4:00AM,
  to be recognized later



Konstantin Todorov
Koen Dercksen

Linguistic post processing

- Problems in OCR can be solved on the input side, but also on the output side

- Goal: to improve garbled digital text output from OCR to something that is closer to the target language

Konstantin Todorov

Deep learning approach

- BERT – state-of-the-art in current NLP tasks
- Transfer learning:
  - Use pre-trained BERT
  - Fine tune so it learns Old Dutch
- Machine translation model
  - Assume uncorrected and corrected as two different languages
  - Train a neural network to transform uncorrected to corrected text

Translation goal:

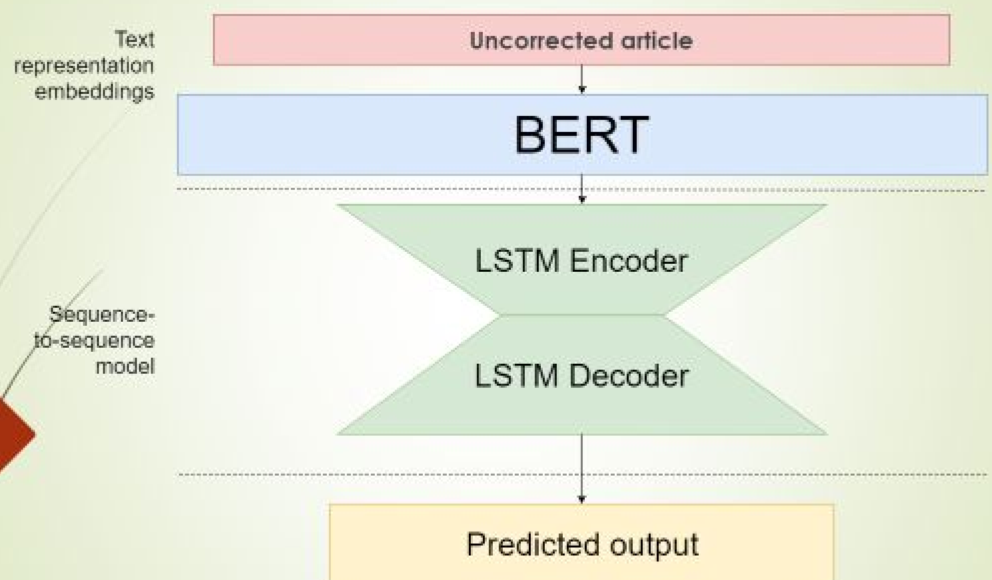
Uncorrected

Uyt der MofÃ¨ou heeftenen van den 14 Juny , dat den gednriocn brandt in die Stadt continueerde, apparent doÃ¶rmoortbrandersaengflftieht, op den n Juny branden alleen 5000 Hiiyfen , vecle Ammunitie-, ende (f 44 menfchcn.opden 11 Juny Â£000 Huyfcn met vyf Czuarfc Cancelycn endeeen groot getal menfcfien

Corrected

Uyt der Moscou heeft men van den 24 Juny, dat den gedurigen brandt in die Stadt continueerde, apparent door moortbranders aengesticht, op den 12 Juny branden alleen 5000 Huysen, veele Ammunitie, ende 644 menschen, op den 22 Juny 6000 Huysen met vyf Czaarse Cancelyen ende een groot getal menschen

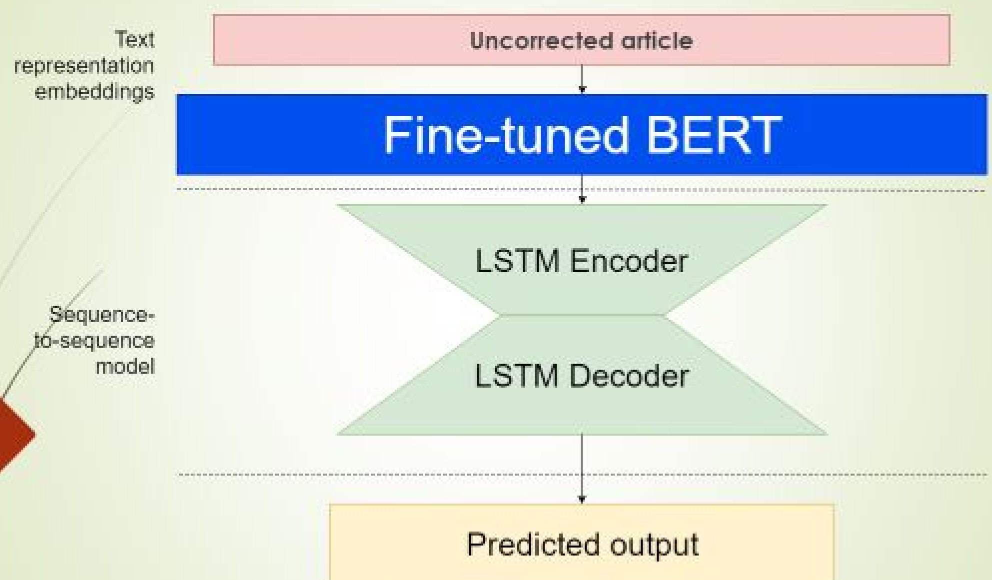
Konstantin Todorov

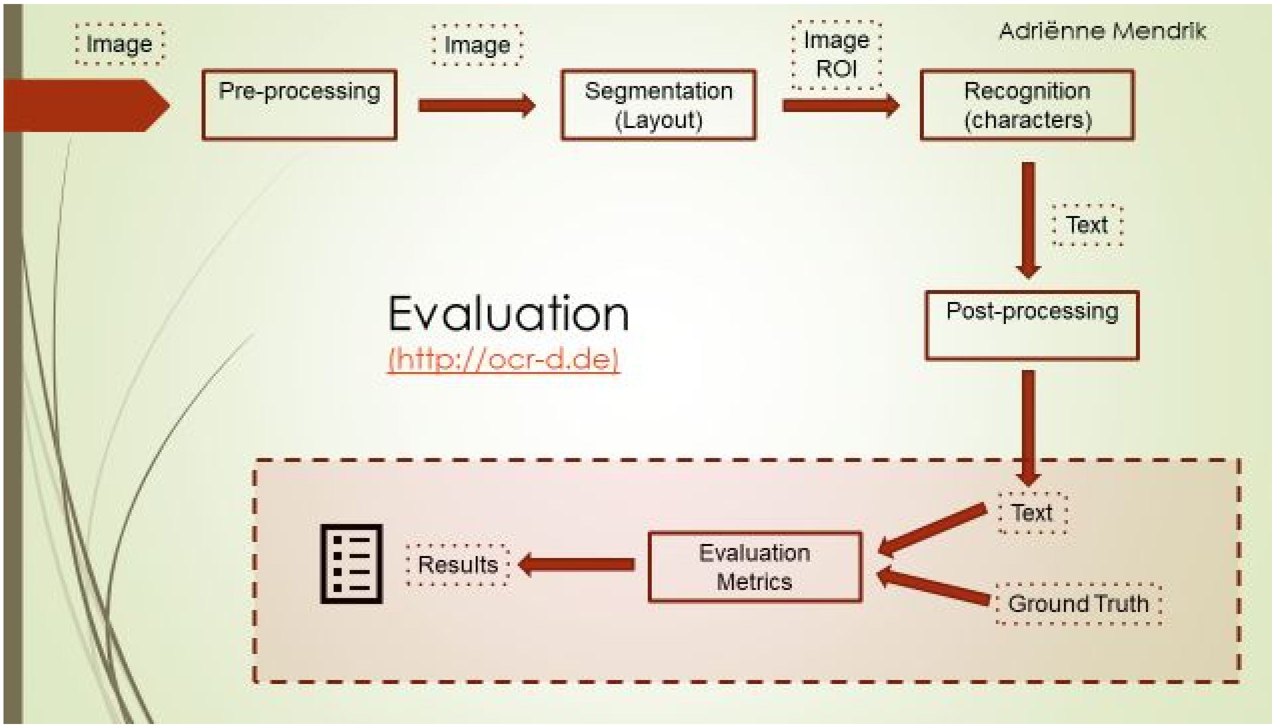


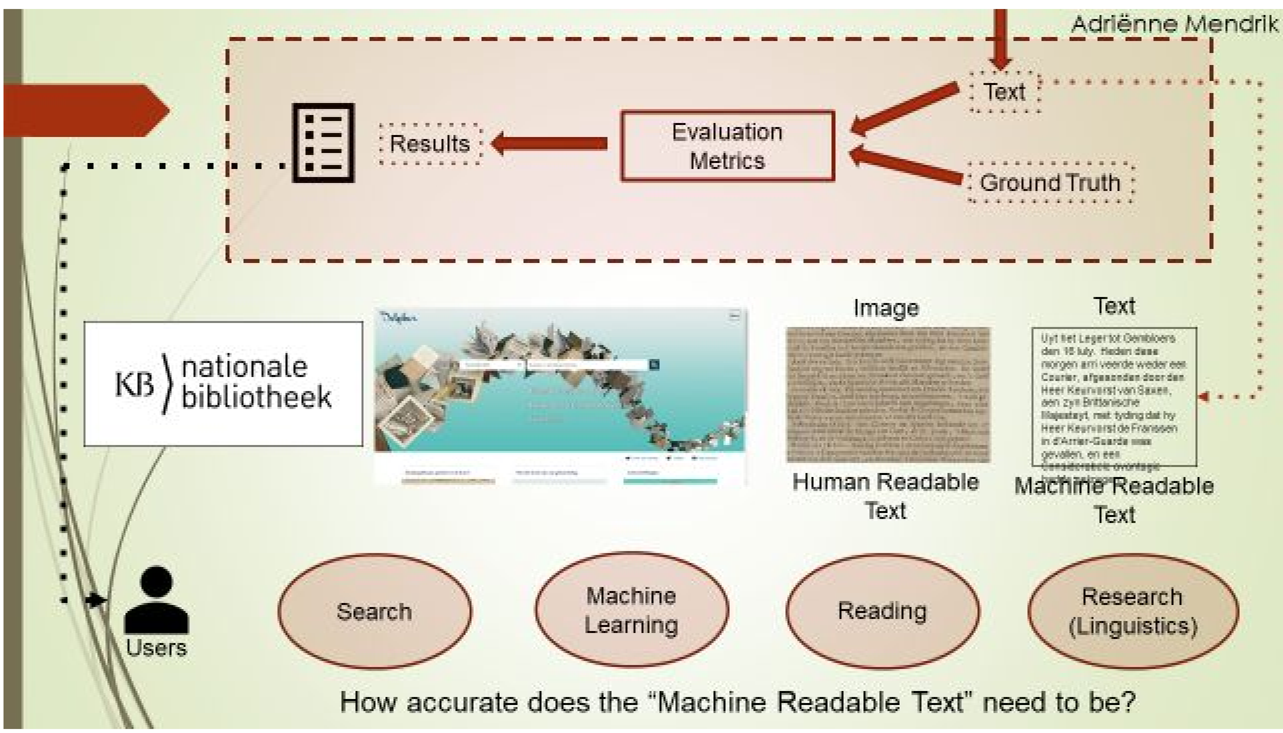
Language modelling

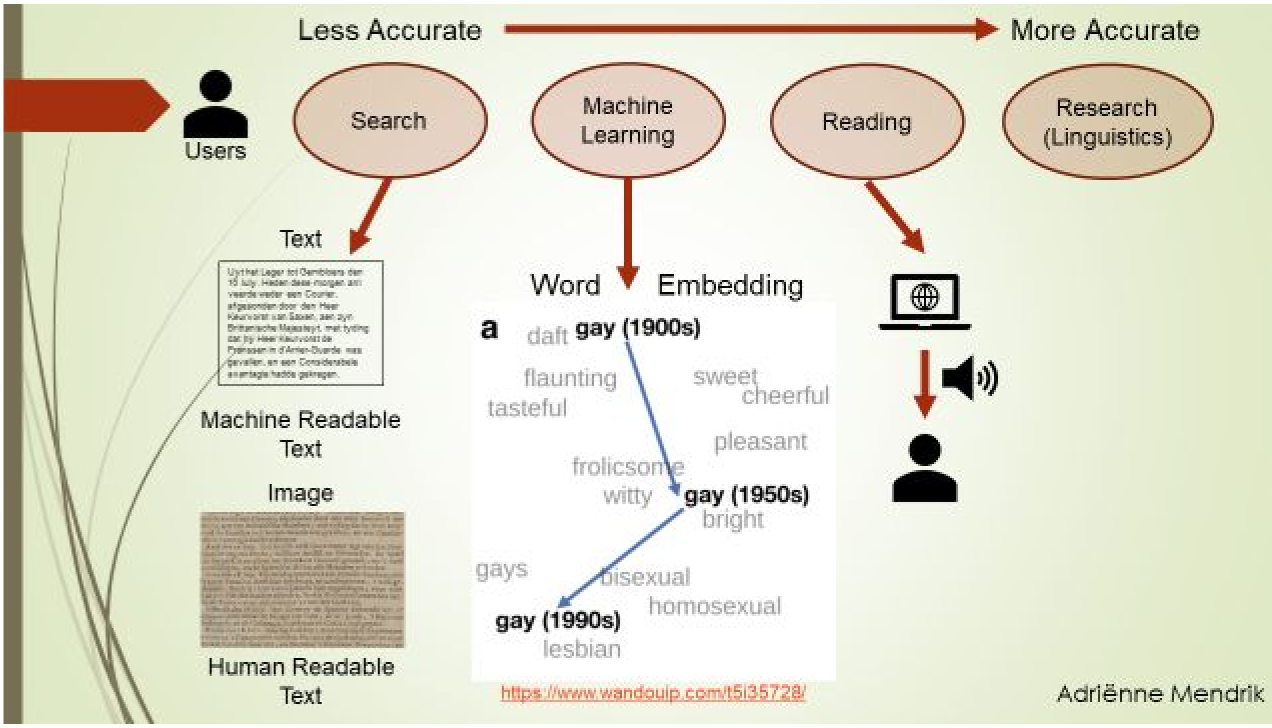
Uyt Venetien, den 8. dito. De Turxse Roovers zijn met sulcken menighte in de **kuste[Golfe]** van Venetien, dat daer **daerlijck[qualijck]** een Schip door passeren kan sonder ghenomen te werden ; **dat[doch]** men hoopt dat daer in wel haest sal versien worden

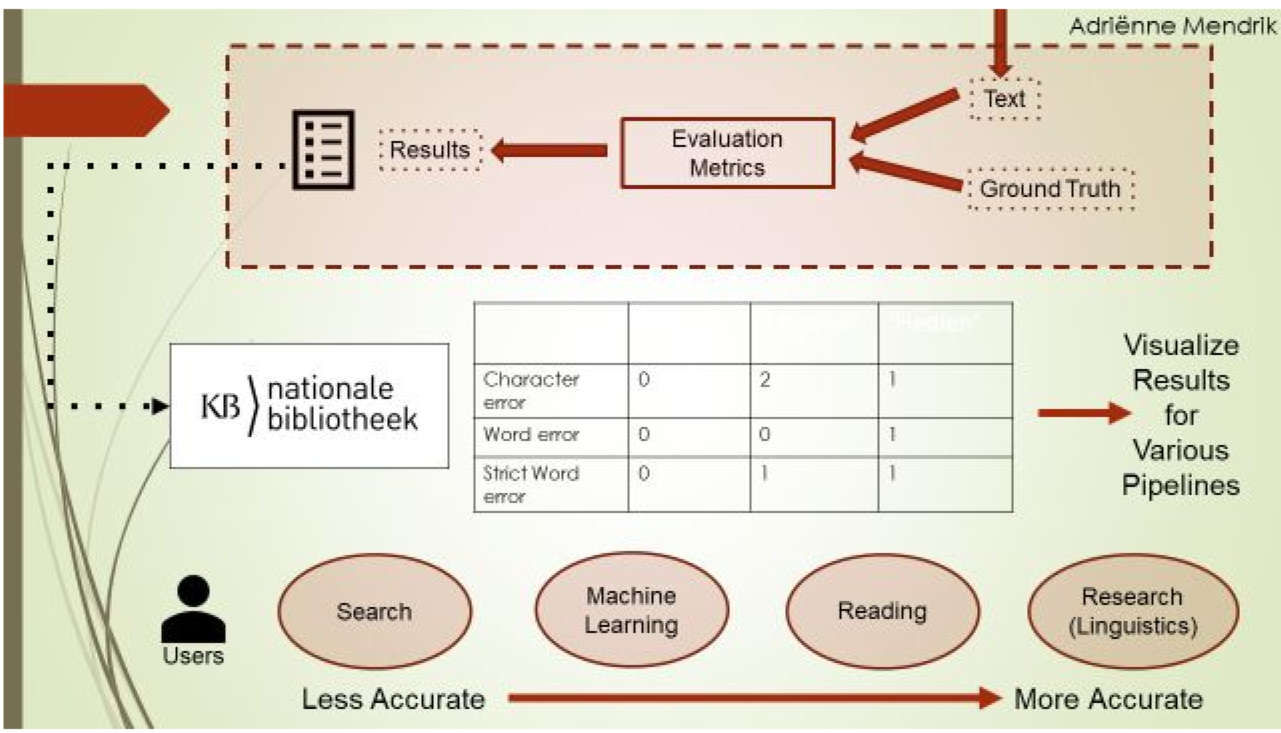
- **Predicted**
- **Original**




Adriënne Mendrik
Image
Pre-processing
Image
Segmentation (Layout)
Image ROI
Recognition (characters)
Text
Post-processing
Evaluation
(http://ocr-d.de)
Text
Results
Evaluation Metrics
Ground Truth


Adriënne Mendrik
Text
Results
Evaluation Metrics
Ground Truth
KB nationale bibliotheek
Image
Human Readable Text
Text
Machine Readable Text
Users
Search
Machine Learning
Reading
Research (Linguistics)
How accurate does the "Machine Readable Text" need to be?



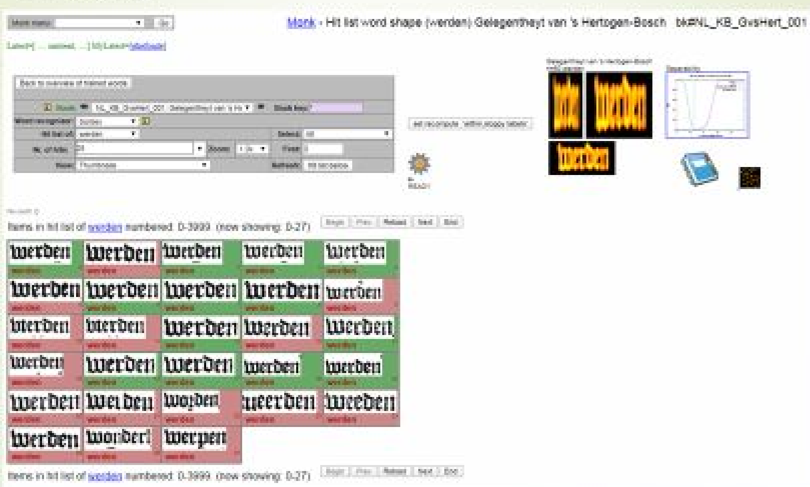


Character error	0	2	1
Word error	0	0	1
Strict Word error	0	1	1

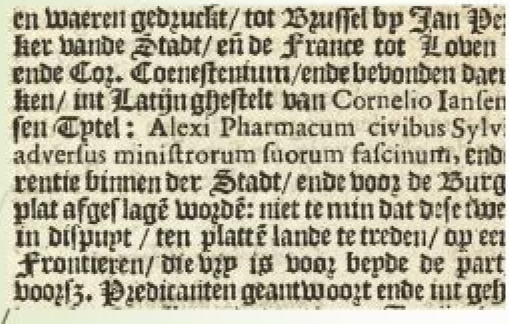
Conclusions /KB

- Monk
- Adjustable OCR Pipeline
- New insights for every OCR step

Monk



The near future… ?



en waeren gedruckt/ tot Bruſſel by Jan Pe
ker vande Stadt/ eñ de France tot Loven
ende Cor. Coeneſtenium/ende bevonden daer
ken/ int Latijn gheſtelt van Cornelio Ianſen
ſen Tytel : Alexi Pharmacum civibus Sylv
adverſus miniſtrorum ſuorum faſcinum, end
rentie binnen der Stadt/ ende voor de Burg
plat afgeſlagẽ wordẽ: niet te min dat deſe twe
in diſpuyt / ten plattẽ lande te treden/ op een
Frontieren/ die vry is voor beyde de part
voorſz. Predicanten geantwoort ende int geh

```
en waeren gedruckt/ tot Bruffel by Ian P
ker vande Stadt/ ende France tot Loven b
ende Cor. Coenestenium/ende bevonden dae
ken/ int Latijn gheftelt van Cornelio Ian
ſen Tytel : Alexi Pharmacum civibus Sylv
adverſus miniftrorum ſuorum faſcinum, en
rentie binnen der Stadt/ ende voor de Bu
plat afgeſlagẽ wordẽ: niet te min dat de
in difpuyt / ten plattẽ lande te treden/
Frontieren/ die vry is voor beyde de par
voorſz. Predicanten geantwoort ende int
```

KB national library of the netherlands

KNAW Humanities Cluster

Delpher

Thanks to the team!