Using Chronicling America's Images to Explore Digitized Newspapers and Imagine Alternative Futures

Elizabeth Lorang, Leen-Kiat Soh, Yi Liu, Chulwoo Pack, and Delaram Rahimi



University of Nebraska-Lincoln projectaida.org



Researchers and developers might do far more with digital images at all stages of digitization and use, and that attention to the digital images will yield greater understanding across a range of domains



Much to learn about the materials themselves as well as about process: (1) values we bring to digitization and those that get enacted through digitization, and (2) how current work extends from those values in maintaining the original items



Studying these processes can function as a form of social and cultural history: how society and culture are captured in collecting and in technology



Exploration and study of digital images of historic materials as a mode of asking questions about the materials of many forms: physical originals, microform duplications, digital copies, represented as images, text, and metadata



What might we learn about digital collections of cultural heritage materials, and how might we augment use and access of these collections, if we focus on the digital images being created as librarians, archivists, museum professionals, and others are digitizing cultural heritage materials?



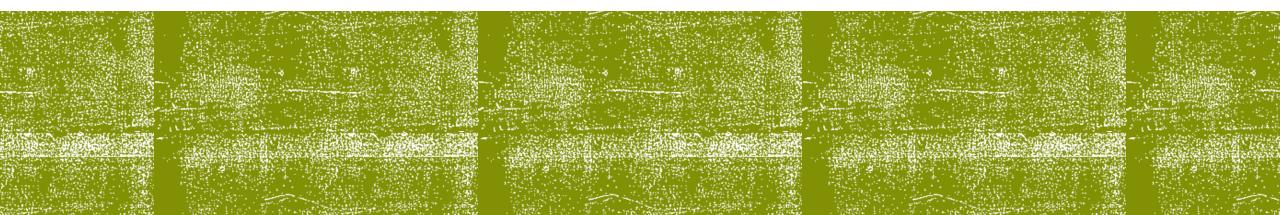
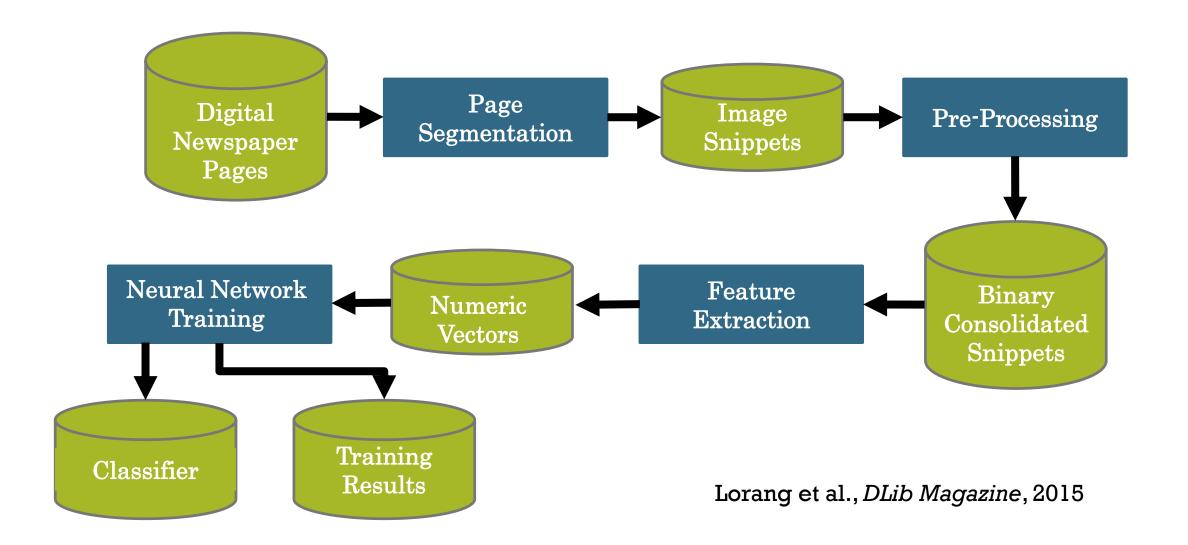


IMAGE PROCESSING

Generate data about visual features from the newspaper pages and then use those extracted features within a computational system called an artificial neural network







CASE STUDY

Test the effectiveness of our first-generation approach on page images from Chronicling America from 1836-1840 (8,000 pages, downloaded in 2016)

The results lagged well behind what we saw in our smaller proof-of-concept project



NOISE

Washington Hall, University of Virginia.

At a called meeting of the Washington Society this evening, Mr. James L. Orr, arose and said, Mr. President it becomes my painful duty to announce to the Society the death of one of our honorary members, Thomas Butler Bird. of South Carolina. When the melancholly intelligence first reached this place, some faint shadow of hope, as to its truth, prevented our giving entire credence to the tragical affair .-But it is now too sadly confirmed, and our much esteemed friend sleeps in death's icy embrace. He has been cut off in the spring time of his existence, and we are left to weep over the many generous qualities of his nature-the bud was just opening-its promised fragrance was adding new charms to its loveliness-but alas! it has been thus early nipped by an untimely frost, and consigned to wither and decay.

"All that's bright must fade, The brightest still the fleetest; All that's sweet was made. But to be lost, when sweetest."

When we reflect that he was distinguished alike for his benevolent spirit, a nobleness of heart, and a superiority of talents, the sympathetic tear starts to swim the eye and moisten the cheek, on account of his unhappy fate. I shall attempt to pronounce no onlogy on his character, but the sorrowed countenances

good quality

Save when some quick emotion.
The warm blood strongly sent.
To revet in her office check, I said consumption smote her. ut she was un Indian meiden So none deplor'd her min;

None, save that widow'd mother,

Who now, by her open tomb.

Is writhing like the smitten wretch

Whom judgment marks for doom. Alast that lowly cabin. That south beside the wall,
That seat beneath the mentling vine,
They're lone and empty allWhat hand shall pluck the fall, green corn. That ripeneth on the plain, Since she, for whom the board was spread, Must ne'er return again? Rost, rest, thou Indian maiden!-Nor let thy murmuring shade (scorn Grieve that those pale-brow'd ones with Thy burial-rite surveye.;—
There's many a king, whose funeral A black-rob'd realm shall see, to be a surveyed to be a surveyed to be a business of grief is shed.

Like that which falls for thes. Tes, rest thee, forest-maiden!
Beneath thy native tree;
The proud may boast their little day,
Then sink to dust like thee; at there's many a one whose funeral

bleed-through

Wit may firger at wait. We mourn a facing vacancy Museber one con fill. Hearts that have lost so dgar a formit Must own one rangest rela-There may in those that they can lave, But more they he love as well. There must be toured purental hopes: Are mediated on death's dark wave, And also who should like a poys have should, Would have have shared his grass ! Food backers moore the first deep great They e'er were documed to know Swiers, whose lave may make be told Are beinling weath the blows And friends-who clong to each foil hope. The base with life min done : And resemble O, why that word? His hard one over one? There inords, sh! how dendate In charm, and half, and bearthy Notice that has from the weed and A gent of principle worth-When from the ga herest worshippers. Acrenda the ferrent program. In coin he seem the benching surgery, One form no more or there, And when the the lling runos join. To ratio the evening song, The increasy is not complete, Our tunidal court is gosto, The friend, so true to make a lengt. The favored one sus flury. To find affective's companies, Robinsons an amoreting tone,

'ANTINE, ST. JOSEPH CO. From the Baltimore Visiter. WE MAY BE HAPPY YET. Ah! dearest dry those tears away, That stain thy fading cheek; Unbend thy lips from sorrows sway, And words of comfort speak. Banish the past, and with me vow Our sorrows to forget ; And be Hope's star our pilot now,-We may be happy yet. The care, believe me, that enshrouds Thy check's once cheerful ray, Gives me more pain than all the clouds That darken o'er our way. Then let thy sweet lips smile again— Smile as when first we met, Grief cannot always shadow them-We may be happy yet. Gaze on you star so bright and clear, Free from its cloudy chain; Thus will our sorrow disappear,

low contrast

occluding "blobs"

When thou dost smile again ;

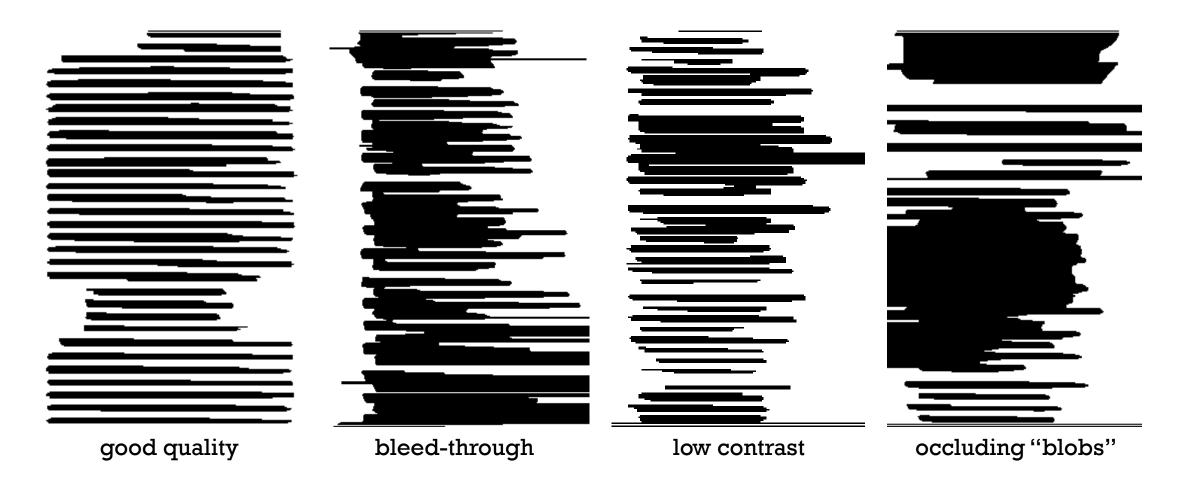


IMPROVED APPROACH

Improve our binarization of the image into "object" and "background" information (represented as black and white pixels, respectively) and extract visual structures with a noise-tolerant text line segmentation approach

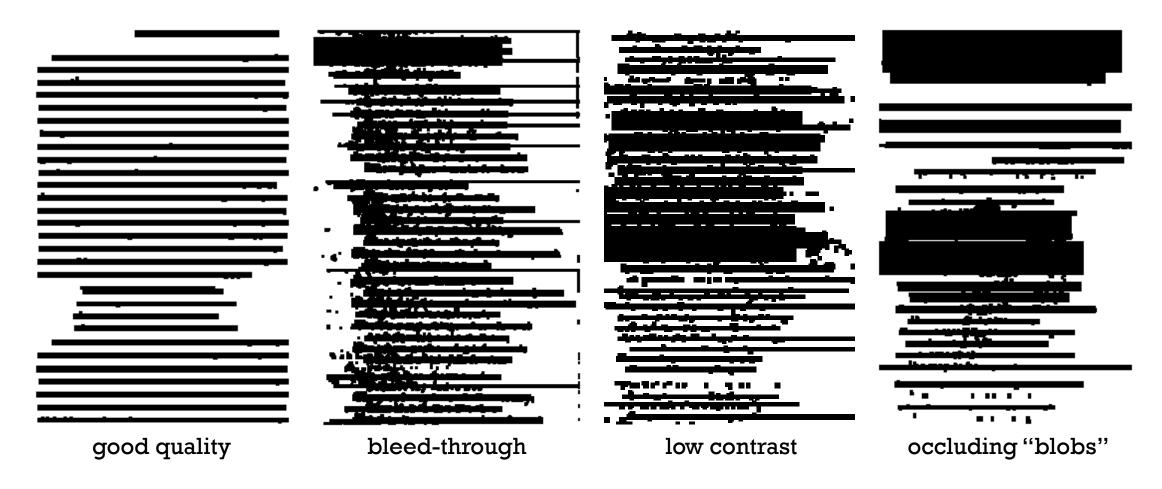


Results from Initial Approach



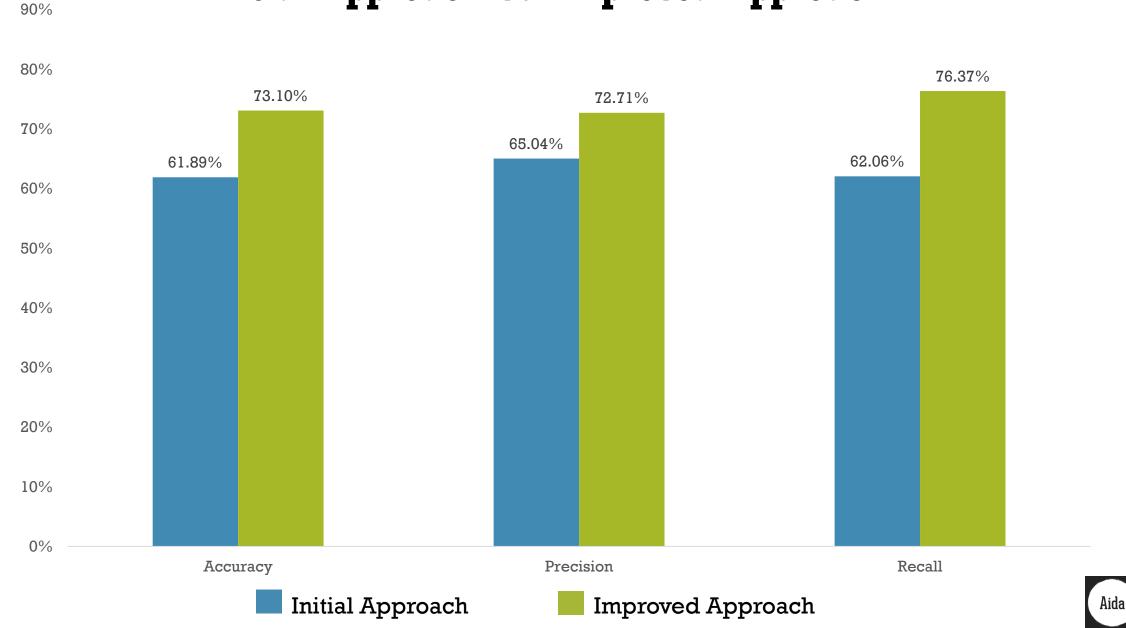


Results from Improved Approach





Initial Approach vs. Improved Approach



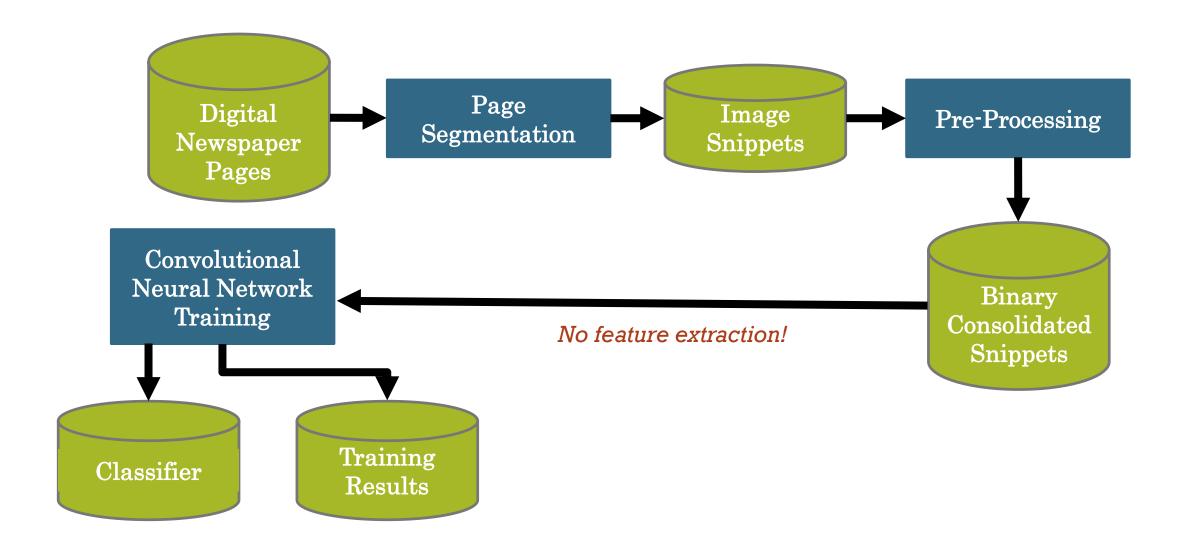
ALTERNATIVE APPROACH

Use deep learning via a convolutional neural network, without having to perform feature extraction at all

Initial results of accuracy achieved 92.4%

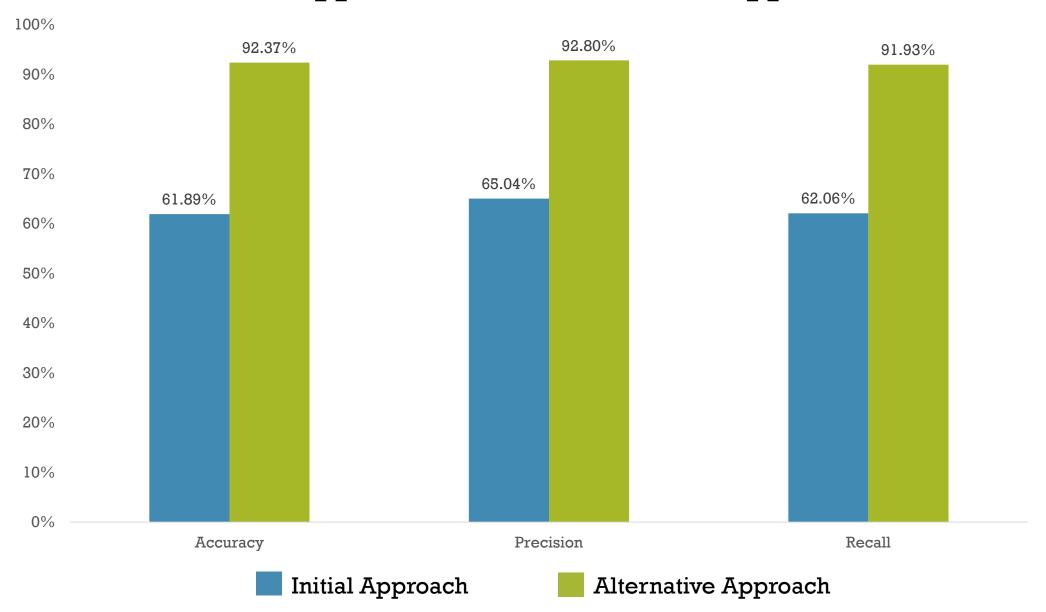
precision = 92.8%; recall = 91.9%



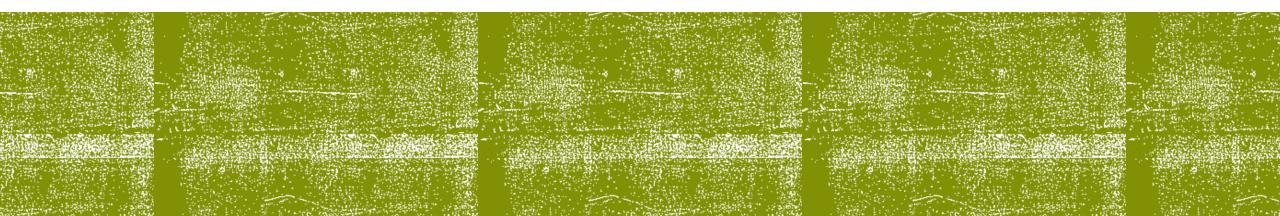




Initial Approach vs. Alternative Approach



Aida LAYOUT ANALYSIS



LAYOUT ANALYSIS

Layout analysis and zoning performed by OCR systems as a preprocessing step for text recognition

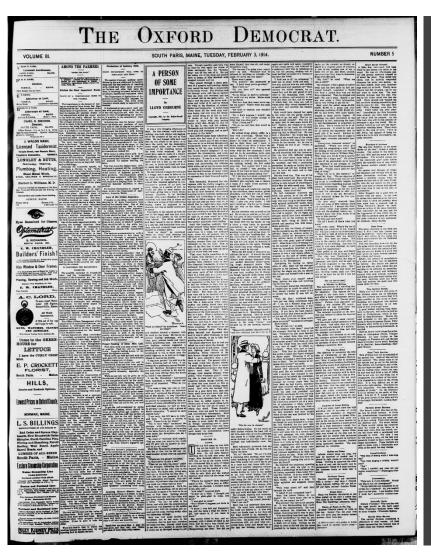
• E.g., built in to OCR web services such as ABBYY Cloud and in the open source Tesseract OCR engine

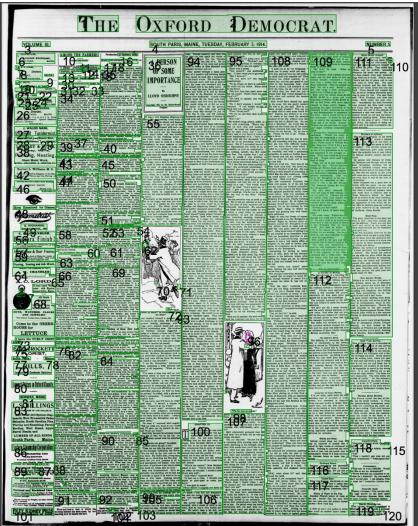
Some of this zoning information is maintained in the OCR XML files made available through Chronicling America:

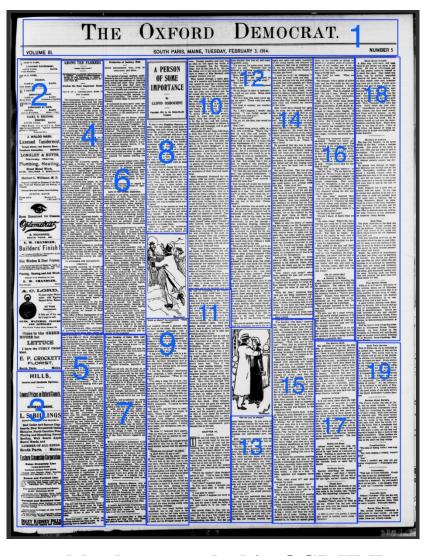
 Zoning coordinates for strings, text lines, some larger text blocks, and page-level information

Layout analysis and zoning remaining challenging for historic newspapers and periodicals, particularly those with dense and mixed layouts







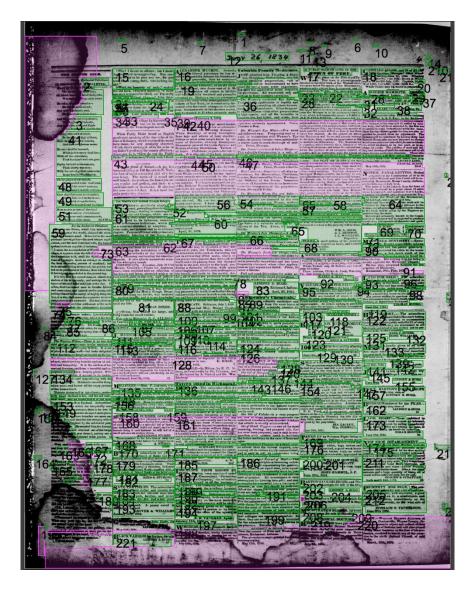


newspaper page from Chronicling America

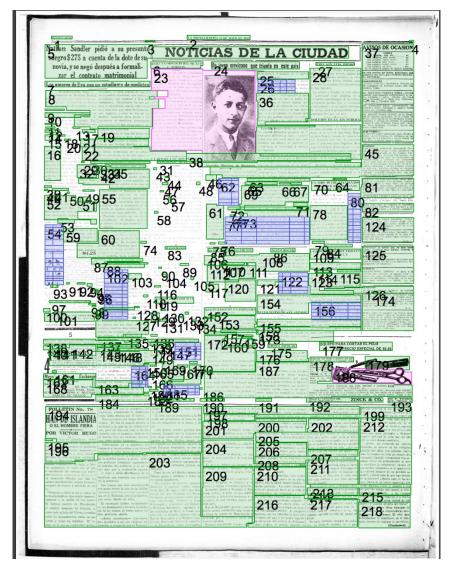
zones identified by Abbyy FineReader

text blocks recorded in OCR XML on Chronicling America





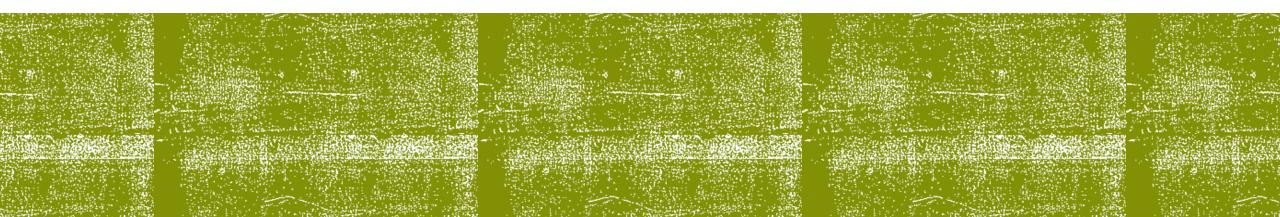
challenges to zoning: damage & noise

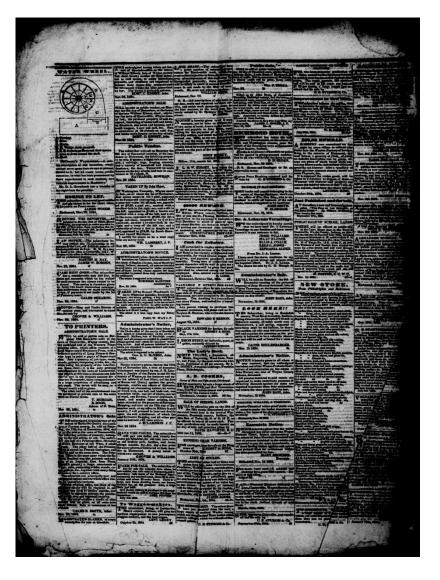


challenges to zoning: contrast and range effect



Aida CORPUS-WIDE ANALYSIS





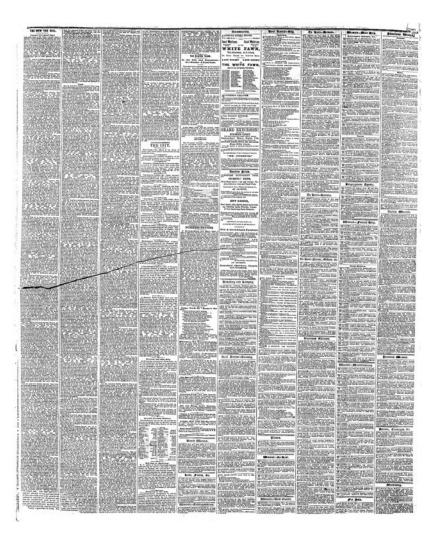


sn82016416_1834-11-29_ed-1_seq-4

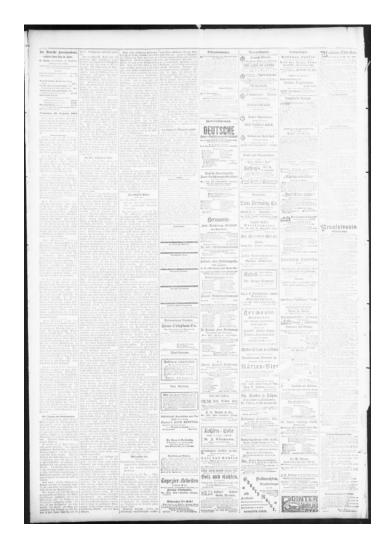
sn90059649_1916-08-04_ed-1_seq-8

 $sn83045348_1871-09-15_ed-1_seq-3$









Newspaper page showing high contrast

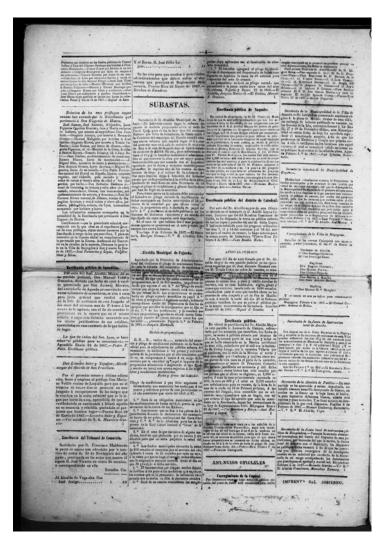
Newspaper page showing average contrast

Newspaper page showing low contrast





high range-effect

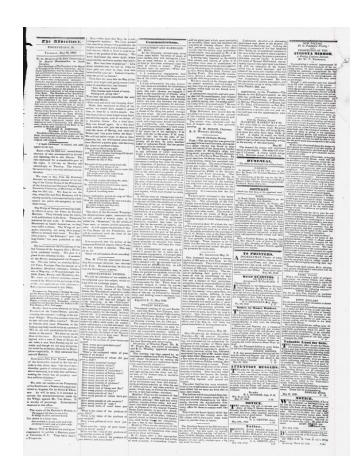


Newspaper page showing average range-effect

og gekk altaf sama vegian. Svo reglubanda - Í Watertown er syddina Jón J. Hoff, er lengi O. G. ANDERSON & CO. ar voru þare ferðir, að mann sögðu það væri bjó hór í grend við Mianeots, niðag í Marshall og O. G. ANDERSON & CO. öhætt að setja klukkur sínar eftir því hvar mi um litinn tíma í Watertown. Jón var ættaði Ibsen væri. Hann gekk altaf niður á sama úr Vopnstirði á Íslandi. Hann lætur eftir sig kom hótelið, settist þar altaf í sama sætið út í horni, og mörg börn, flest uppkomin. Dar las hann morgunblöðin eina klukkustund, unn Gunnarson, kona Páls Gun hvorki lengur né skemur, og tæmdi á þeim var gáfukona og skörungur mikill og var bán við-Sičan Ibsen fékk slagið befir hann aldre sjå hann nema þegar hann sat við gluggann Vér höfum til sölu við vægu verði og fimtán starfsmenn eru í búðinni og skal revnt sama tima dags; og svo reglubundinn var så sem aldrei áður höfðu séð hann gátu hægl O. G. Anderson & Co. N. Dakota. Verð frá \$5.00 til \$15,00 eins á jólum og nyári voru honum ávalt send NY VERZLUN er ÁRNI B. GISLASON, Washburn, Hosis Thorlaksso Frétta Pistlar. bréflega til undirritaðs ráðsmanns félags-HÉSGAGNA-VERZLUN "GLOBE"-FÉLAGSINS. og öskar eftir viðskiftum landa sinna. Bjorn B Gislason, Hann befir einnig íslenzka MINNEOTA MINN helts mi, en tölu vert frost oftast mer. var framin innbrots pjöfnaður á Mountain. Báður Drs. Brandson & Bell, og útvegar allar falenskar bækur, sem út Dr. O. Bjornson, Hósías Thorláksson. land sitt vestur i Mouse River-dal, og býst við að dvelja par vestra un lengri oba skemmri tima. | WINNIPEG - MANITOBA. B. G. Skulnson. Deir herrar Sigurrín og Hermana Jóseps Dr. O. Stephensen, Skulason & Skulason að sjá sig um og heimsækja vini sína. Þeir föru til Winnipeg héðan og halds siðan helm leiðis. John G. Kristjánsson hefir flutt í hús sitt á I. V. Leifur. WINNIPEG - MANITOBA. Clifford Building, GRAND FORKS, N.D. MINNESTA, MINN., DES. 1904. Dr. G. J. Gislason. Bjorn B. Gislason, son, sonur Jensar Sigurðssonar bónda þar í hygð- Weilington Block - Grand Forks, N. J Bes Augnalerningen Vertt St mjög lengi af margskonar sjökdömum. Hann var MINNNEOTA, · · MINNESOTA.

Newspaper page showing no range-effect

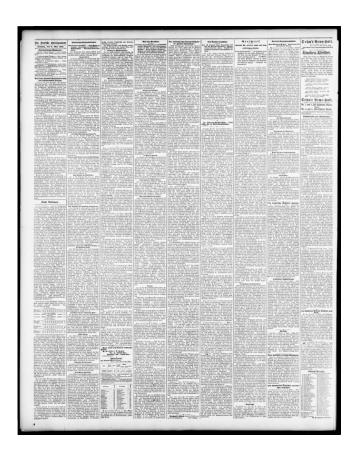




Newspaper page showing severe orientation skew



Newspaper page showing some orientation skew



Newspaper page showing no orientation skew







Newspaper page showing severe bleed-through

Newspaper page showing some bleed-through

VÍNLAND þá framfara tilraunum og veki sundrung og hætta að svo stöddu. Þeir sjá hvað sá öfriður nægju meðal Filippinga, er geti leitt til mundi kosta, og enda óvist hvort beir gætu hann allta poirra gripa felmetra, er hann hafði þens, að þeir vilji segig skilið við Bandarítnin með öllum þeim að er þeir hafa sú, að á sér boriði geillir hans, brjótnall sett der I annan stad er pad audsrett ad Japanar betra verdi ad bida, því ad audsætt er pad, ad hann hatch hatt vrobskitt vid og saged forseta pess raunir sinar. Kaupmaðurinn fór með hann á kirfstefn korreslannar og saged halda Filippseyjum. Ef Japanar væru nú halda Filippseyjum. Ef Japanar væru nú num á skrifstofu lögreglunnar og sagði hanta ritippneyjum. na en sagni nú mest eftir lýðæsingamúnnum í Japan til vel við búnir og treystu sér fyllilega til að nú mest eftir lýðæsingamúnnum í Japan til haft) kann þá und ste ársð, brjóstnáltas og flut strendin skull teðurastreiðis. Það bogasta kan þeim nokkura sinn á átt árkvanstreiðis. Það bogasta kan það og flut strendin strendin skull stær verðiðið þvíða, pratigans gærði hann ekkir skull stær verðiðið þvíða, pratigans gærði hann ekkir skull stær skull s så þá í hvert čefni var komið og þorði ekki að heimta peningana, þó hann vissi vel að þeir væru í vörslum lögreglunnar eins og munir heim veru í vörslum lögreglunnar eins og munir heim veru í vörslum lögreglunnar eins og munir heim veru sö nökkuð eðil landsiss, og þótt sá vegur sö orn, og að líkindum yrði einnig anðveldara þótt landið sé hér milt og anðvelt i sam fyrir þá, en nokkra aðra þjóð, að menta Fil. við óbygðar auðnir, þegar innar dregur í landið, pinga og útrýma þeim villiþjóðabrag, sem þá höfum vér þó kynst að nokkru eðlisfari þessar-m ríkir með ýmsum þjóðflokkum þar í landi. ar mikin eyjar. Vér höfum séð smiðjur elds og Filippingar og Japanar. réttast gert af Bandaríkjum að sloppa hendi Þeir hafa kynst þeim og átt allmikil viðekifti hafa gert þetta land. Vér höfum séð hraunlæstð af Filippingum og selja eyjarsar Japönum. við þá hin síðustu árin, því að margir Japanar ur og hrausspeungur og eldfjóll (8kjaldbreiði) í

Til pessa hafa eyjar pessar verið Bandamönn-um byrðarauki, afðan þeir eigauðust þær, og lit við þá að eigu, telja þá ásælna og sér-sið ett hlð stærsta stöðurata á Íslandi og fyrir hafa, að söga kunnugra manna, kostað þá dræga og gruna þá jafnan um svík og pretti.
freklega 200 milj. dollara, og enn eru litlar En þegar tortrygni er vakin með hálfment- um engjum og stérdjótum. Þó hötum vér eigi só likur ill að Bandaríkin fái nokkuru sinni það aðri þjóð, eins og Filippingar eru, loðir hún nema lition hluta ef hinu mikla landi, þar sem sjá fé borgað með tekjum af eyjunum. Hins veg- lengi við, og verður seint upprætt til fulls. mi hinn hrikalegssta miklifengleik eðlisafanna, ar litur út fyrir, að þau haf enn lengi af þeim Þess vegna eru Filippingar Japönum mjög regar lan á eyðinserkornar kemur. Ísland er mrinn kostnað, þar sem þau verða að halda fráhverfir nú sem atendur, og mandu einski, "undarlegt sambland af frosti og funn" eins og par uppi landvörn, hafa herskip par við eyj-síður óska en þess, að komast undir þeirra skáldið kvað og kvergi á jörð er seitt annað land arnar, og kastala á landi, sem þau sjálf við-vfirráð. Auðvitað eru þeir engir mena til að ser sem eldi og és er blandað samas með jafnfurð habla a eigin kostnað. En Filippingar eru mæta Japönum í stríði, en þó má ætla að þeir seinir til framfara og geta naumast montast gætu unnið þeim mikið ógagn og orðið nokkra, sem ganga fram úr hinum miklu ishreið svo 4. cinum mannsaldri, að þeir verði færir Bandamönnum mjög að liði, ef Japanar kæmu um, er rekja bálendi Íslands og teygja skriðjúkla til að sjá um sig sjálfir, svo að óhætt verði að þangað með ófriði og vildu taka eyjarnar og jökulelfæreiður um dali og slóttur. Viðista

sleppa við þá fullri sjálfstjórn. Ekki er þess heldur að vænta, að Bandamenn geti fyrst um Dó Bandamenn hafi enn ekkert annað en jökull einn er 150 fermilur^a). po na o merkedir menn poon ink ioristagel par i um, på vitja þorr tyrer nivera mun handa pam., resucen vie ag ve soms stor gu og færsb. inter governe viera menn handa pam., resucen vie ag ve soms stor gu og færsb. inter je store i pam. På pam. På

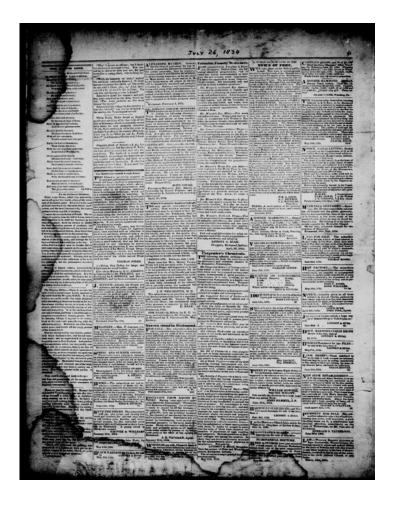
sinn átt arðmiklar nylendur þar í eyjunum, erfiði og kostnað af yfirráðunum á Filippseyj.

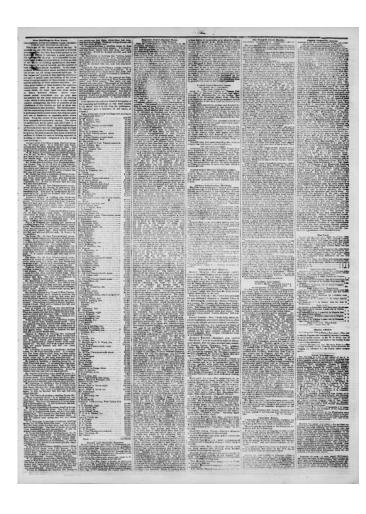
Viða blenda éguring eldfjöll undir ísfærgi
því að hérlendir menn þola illa leftslagið þar um, þá vilja þeir fyrir hvern mun halda þeim.

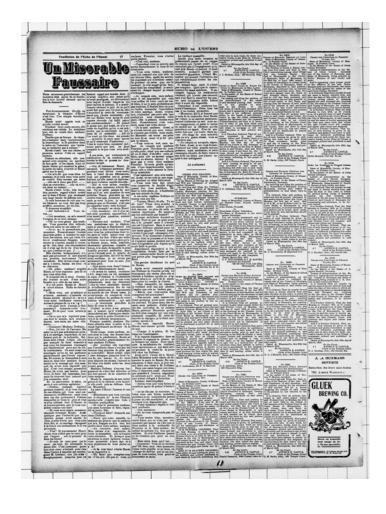
Jessu; en við eg við koma stór gos og bæða íslus

Newspaper page showing no bleed-through









Newspaper page showing severe noisiness

Newspaper page showing some noisiness

Newspaper page showing no noisiness



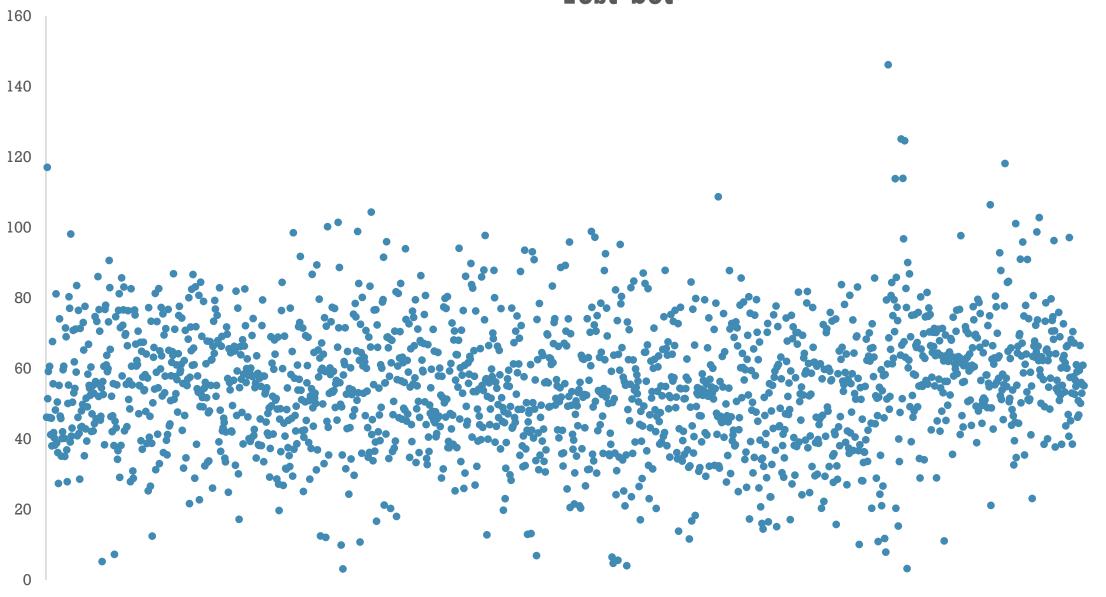
EVALUATION FEATURES

Calculated and analyzed each page and grouped results by language for several features:

- Contrast
- Range effect
- Orientation skew
- Noise

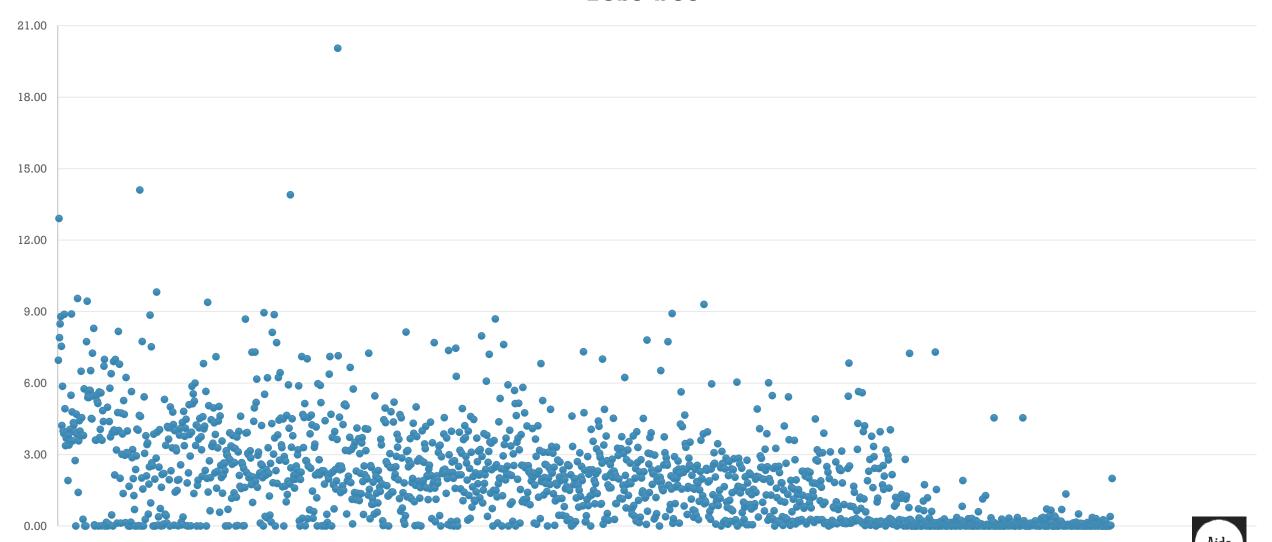


Contrast Test Set





Range Effect Test Set



ANALYSIS: CONTRAST, RANGE EFFECT

Most newspapers have contrast values between 40 and 80, with contrast > 30 demonstrating good contrast

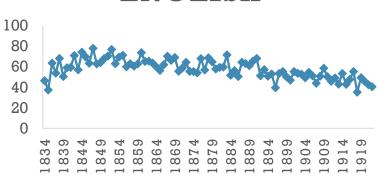
- The lower the contrast value, the worse the contrast
- ~ 6-7% images have bad contrast

Range effect becomes challenging when > 3

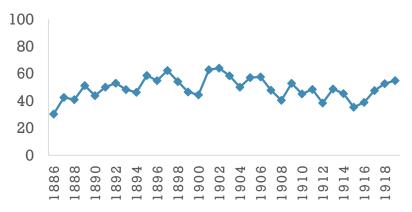
- The ideal value for range effect is 0 (= no range effect is present)
- ~ 25% images have challenging range effect



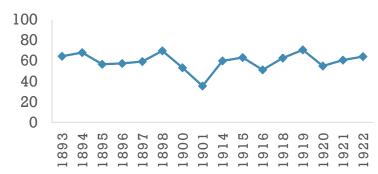
CONTRAST: ENGLISH



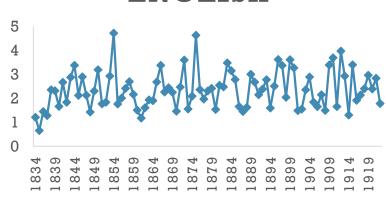
CONTRAST: POLISH



CONTRAST: NORWEGIAN



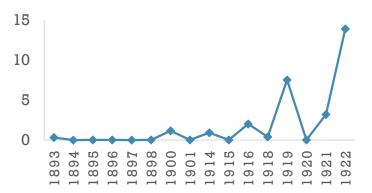
RANGE EFFECT: **ENGLISH**



RANGE EFFECT:



RANGE EFFECT: **NORWEGIAN**



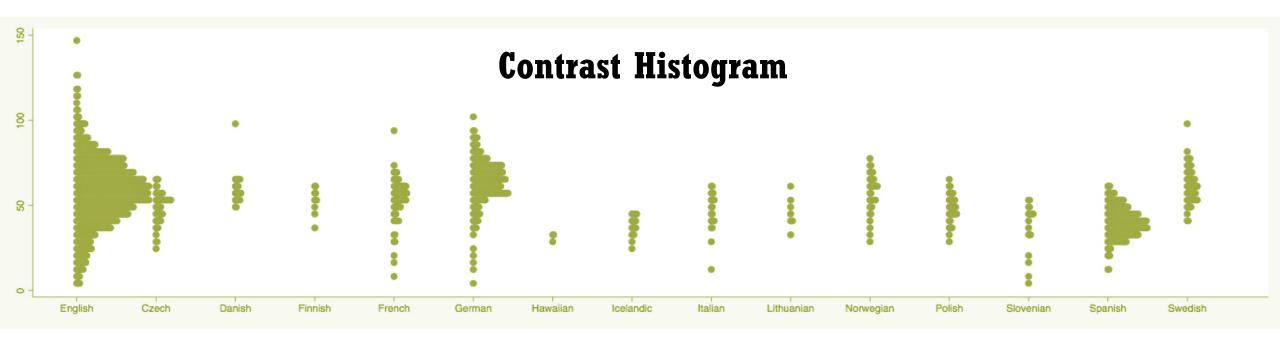


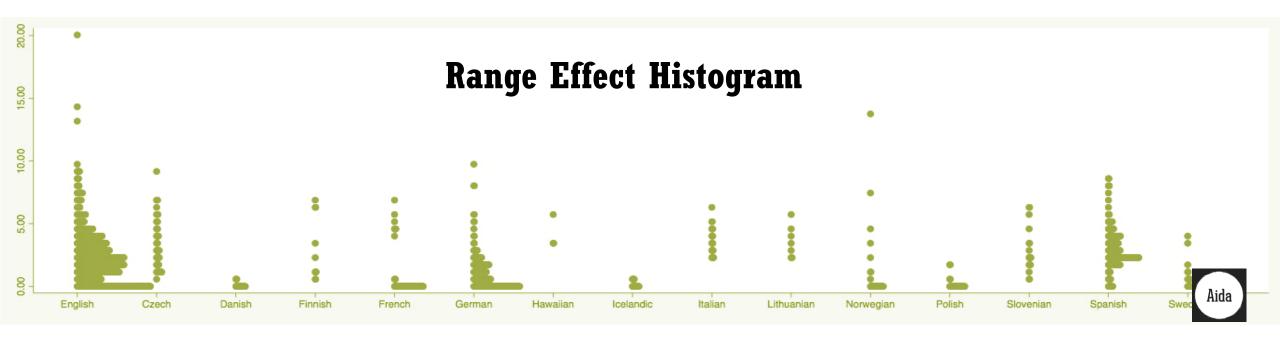
ANALYSIS 2: CONTRAST, RANGE EFFECT

Contrast in all languages is pretty consistent; nor does it change drastically over time

Range effect, on the other hand, not only varies across the different languages, it also changes over time for each language







ANALYSIS 3: CONTRAST, RANGE EFFECT

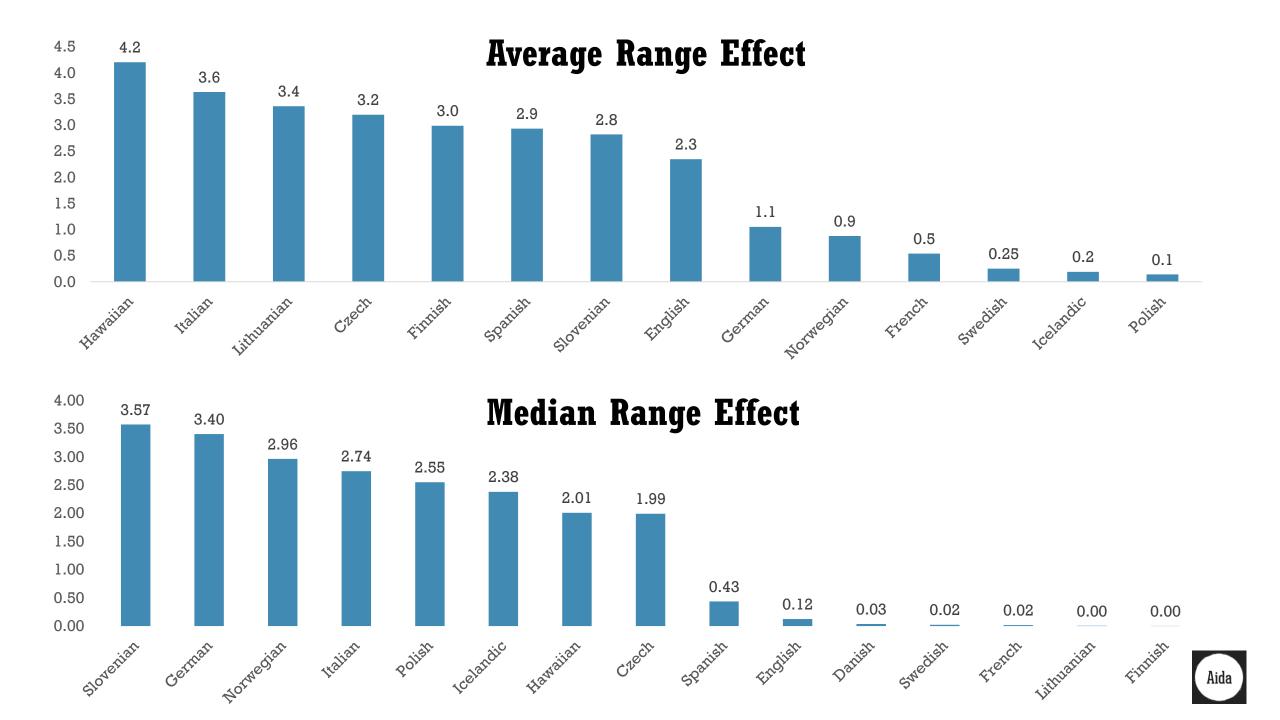
From the histograms:

The data on Contrast are fairly symmetric, while the Range Effect data are right-skewed

A lot of outliers in our Range effect data

•The presence of outliers in our range effect data indicate that the mean (average) would not provide a good estimate: instead, median would give us a good estimate to evaluate the center of the data





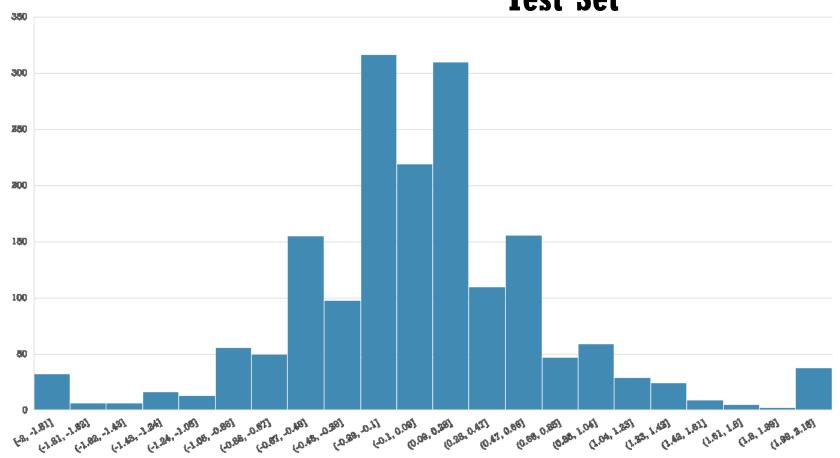
ANALYSIS: ORIENTATION SKEW

Global orientation skew of the pages in the set

A more effective measure is likely to be local skew, relative to particular parts of the page, or other measures of warpedness or beveled nature of the page.

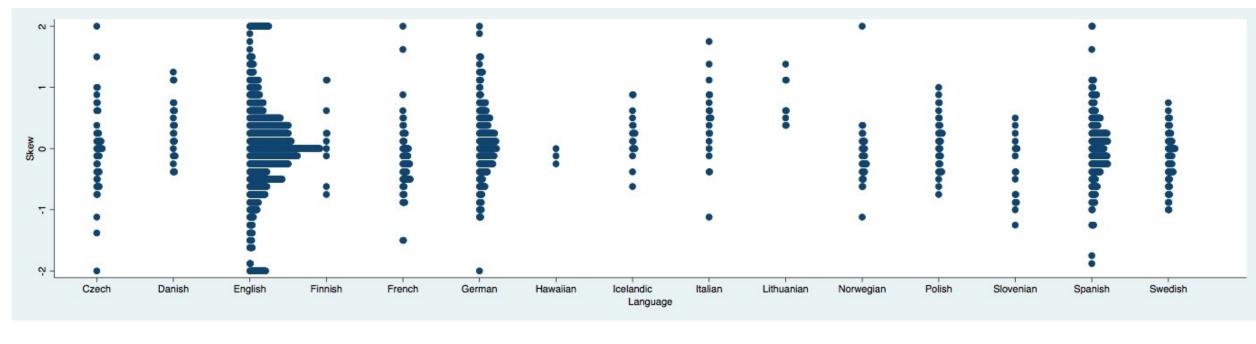


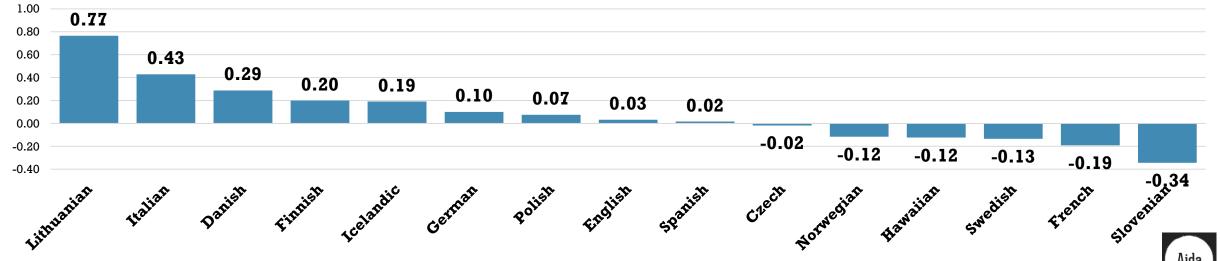
Distribution of Orientation Skew Test Set



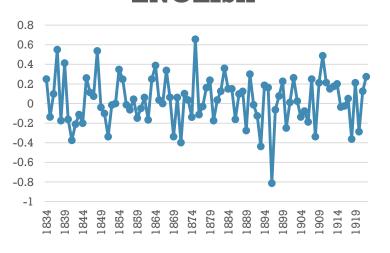


Views of Orientation Skew

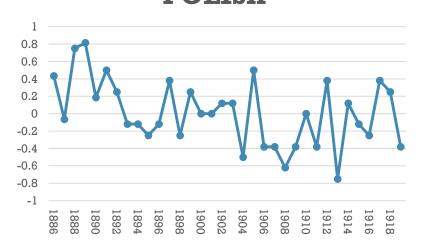




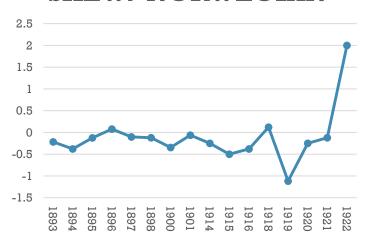
ORIENTATION SKEW: ENGLISH



ORIENTATION SKEW: POLISH



ORIENTATION SKEW: NORWEGIAN









The forgotten labour of digitisation. All newspapers in the British Library were ironed before they were microfilmed to ensure a clear image. Here is a staff member ironing a paper. The microfilm images are used in modern digital packages. #s482 #KZoo2018



2:01 PM - 12 May 2018



ANALYSIS: NOISINESS

Assessing effects of bleed through, blobs (e.g., stains), and other non-textual artifacts

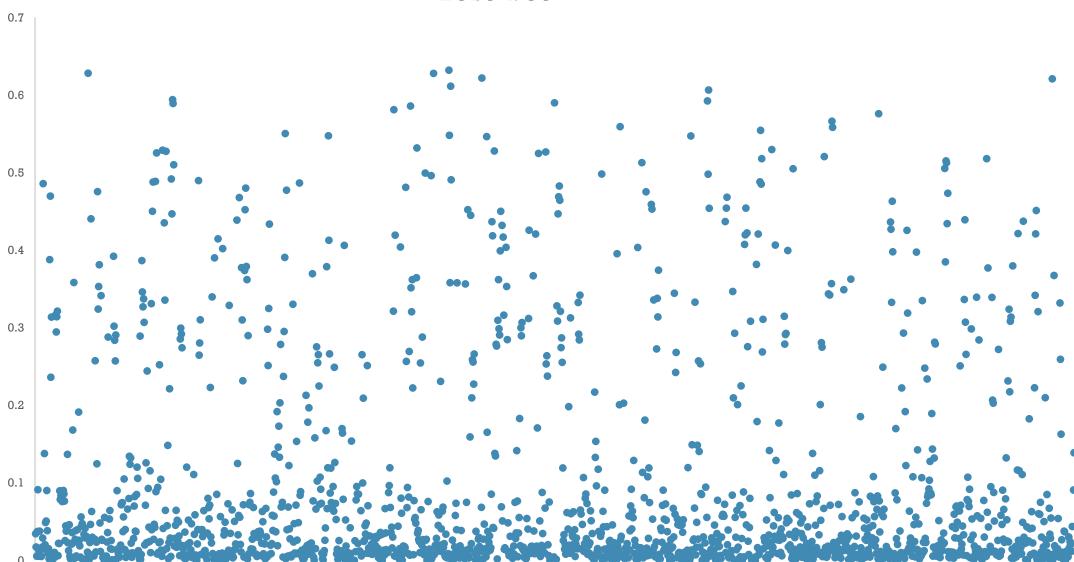
Defects or degradations of a page, or of the digitization process
 Based on histogram analysis—of pixels' intensity values—of each page

Our current technique:

- If an image is deemed noisy, then it is indeed noisy
- If an image is deemed NOT noisy or of low noise then it is either (1) indeed NOT noisy or of low noise, or (2) of poor contrast

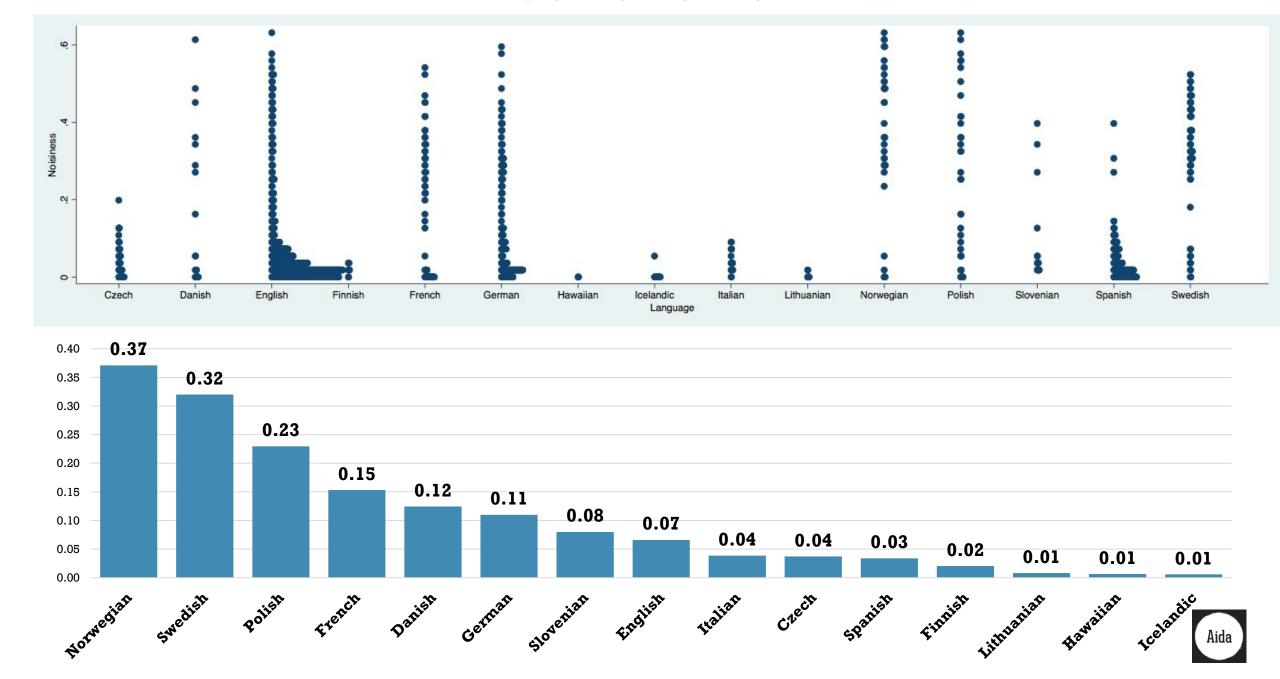


Noisiness Test Set

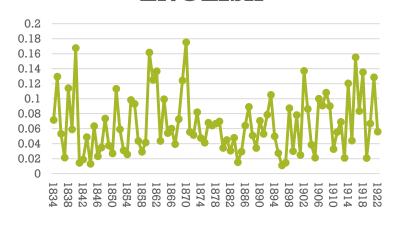




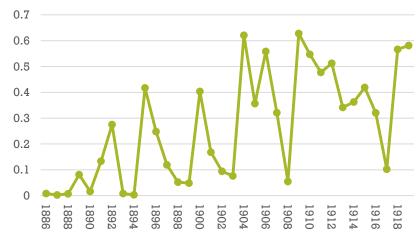
Views of Noisiness



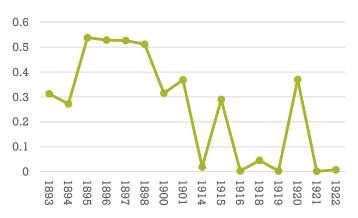
NOISE EFFECT: ENGLISH



NOISE EFFECT: POLISH



NOISE EFFECT: NORWEGIAN





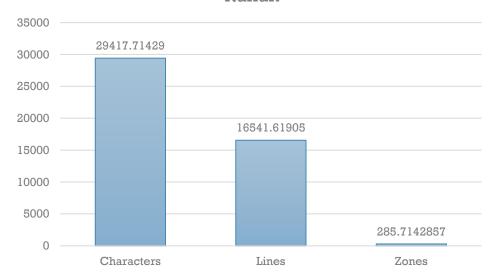
ANALYSIS: CHARACTERS, LINES, ZONES

Compute the number of characters, lines, and zones using zoning technique

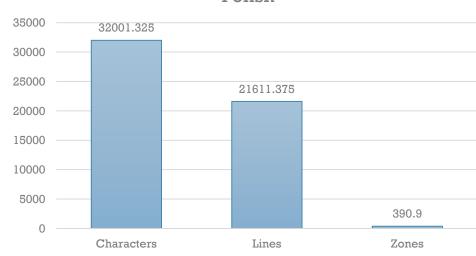
As a way to begin measuring both complexity and density of the textual content on the page



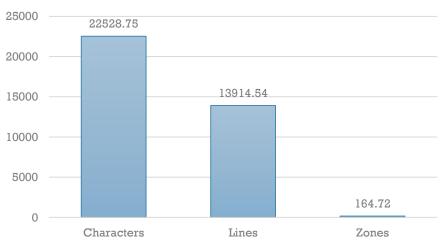
Italian



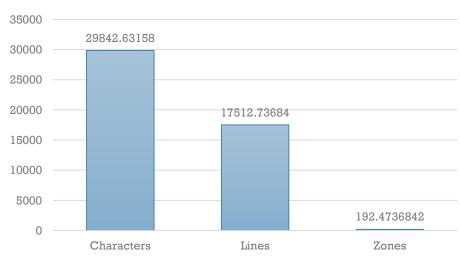
Polish



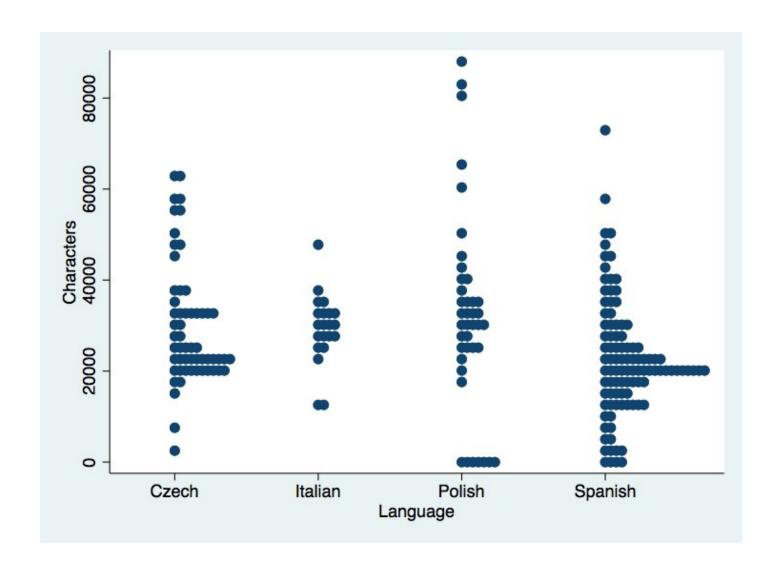
Spanish



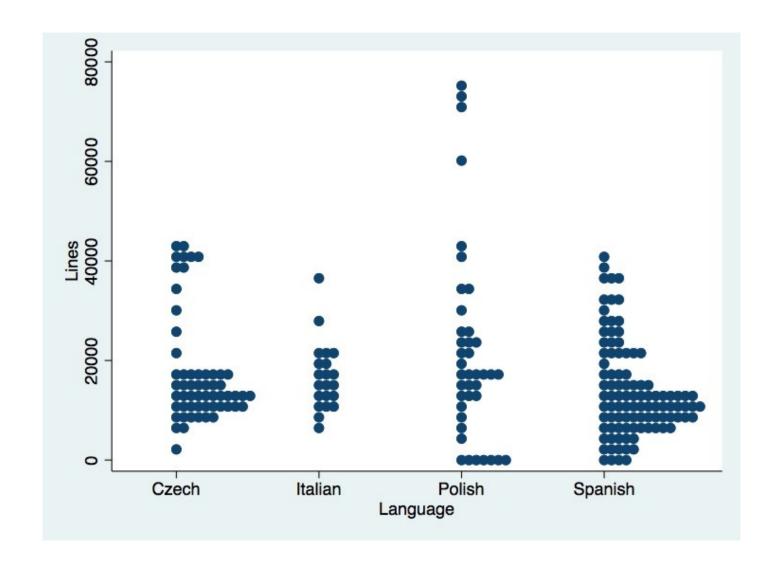
Czech



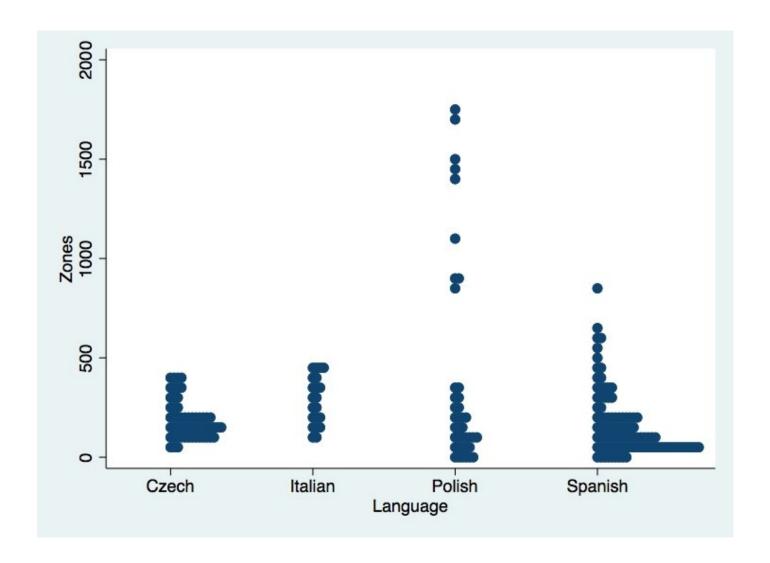














ANALYSIS 2: CHARACTERS, LINES, ZONES

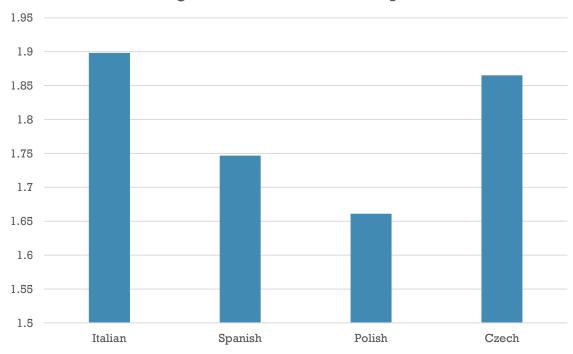
Numbers of characters, lines, and zones vary across newspapers of different languages

- Polish, Italian, Czech: more characters per page than Spanish
- Italian, Polish: more zones per page than Spanish, Czech

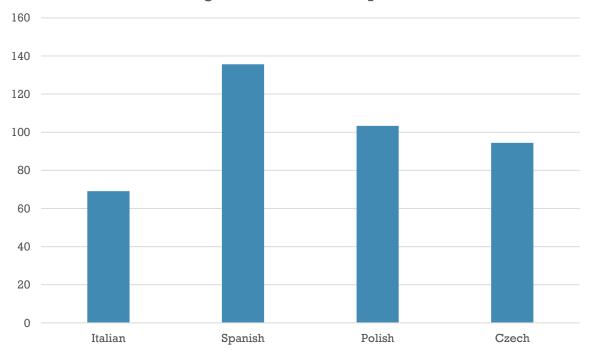
Polish-language newspapers had more outliers; Italian-language newspapers more consistent



Average number of characters per line



Average number of lines per zone





ANALYSIS 3: CHARACTERS, LINES, ZONES

Italian- and Czech-language newspapers had more characters per line

- Does it mean they had more compact typesetting?
- To investigate further: average length of lines

Spanish-language newspapers had more lines per zone

- Does it mean Spanish-language newspapers had more uniform layouts such that a detected zone was not broken up into multiple smaller zones?
- To investigate further: average size of zones

