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

Aida

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

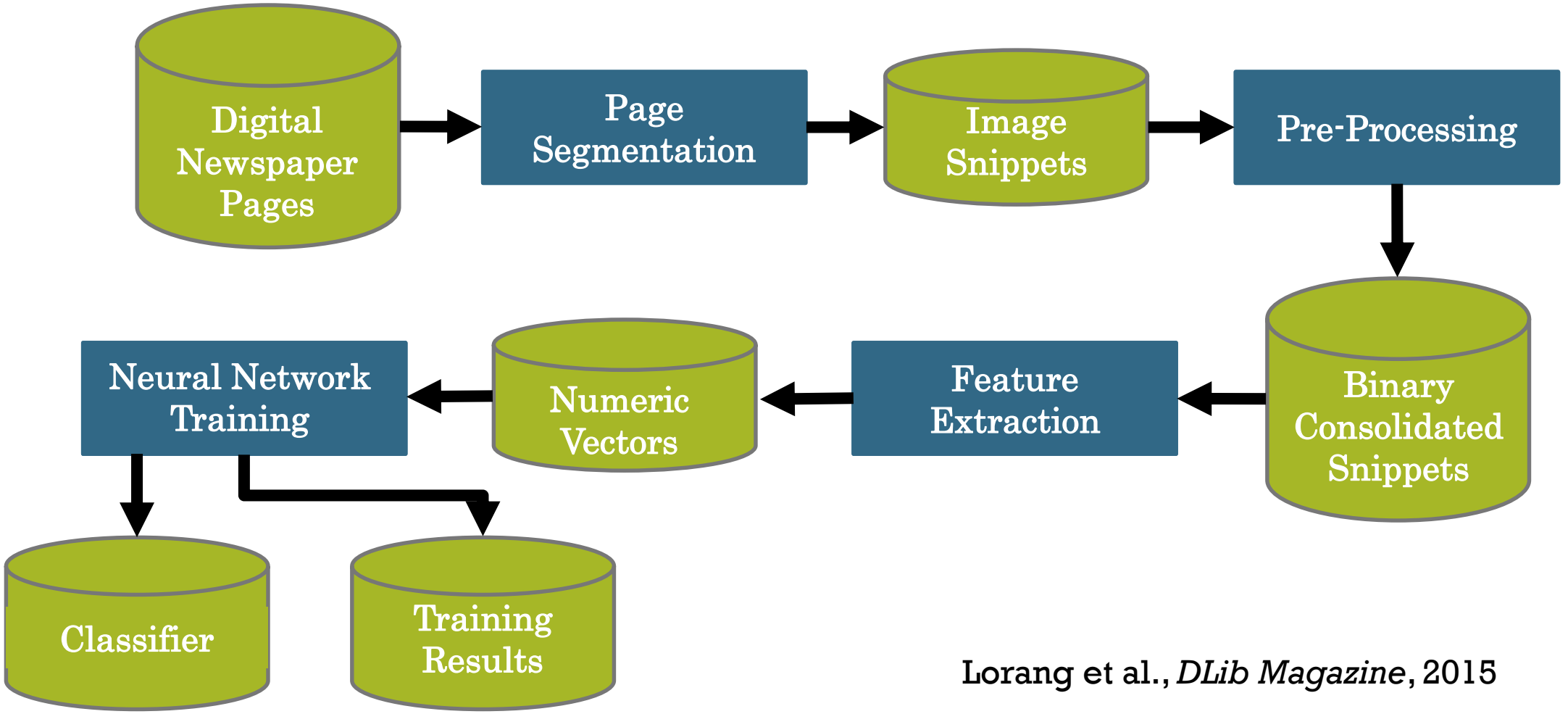
What might we learn about digital collections of cultural heritage materials, and how might we augment use and access of these collections, if we focus on the digital images being created as librarians, archivists, museum professionals, and others are digitizing cultural heritage materials?

POETRY

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

Aida



Lorang et al., *DLib Magazine*, 2015

Aida

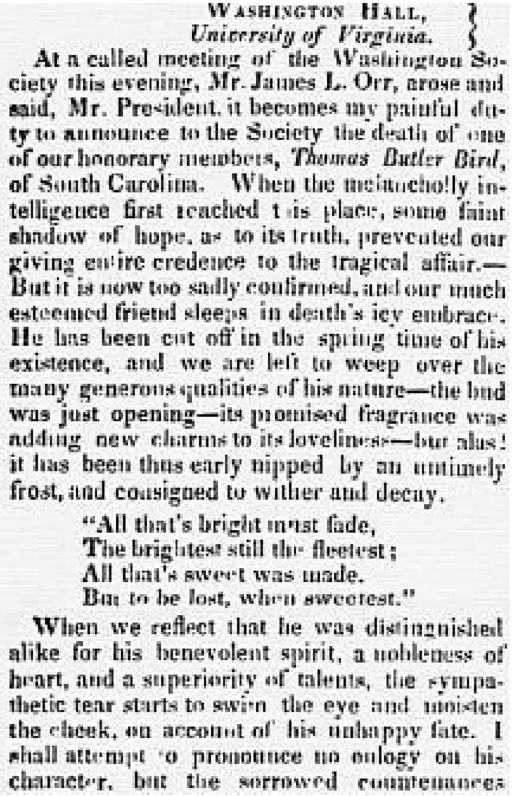
CASE STUDY

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

Aida

NOISE



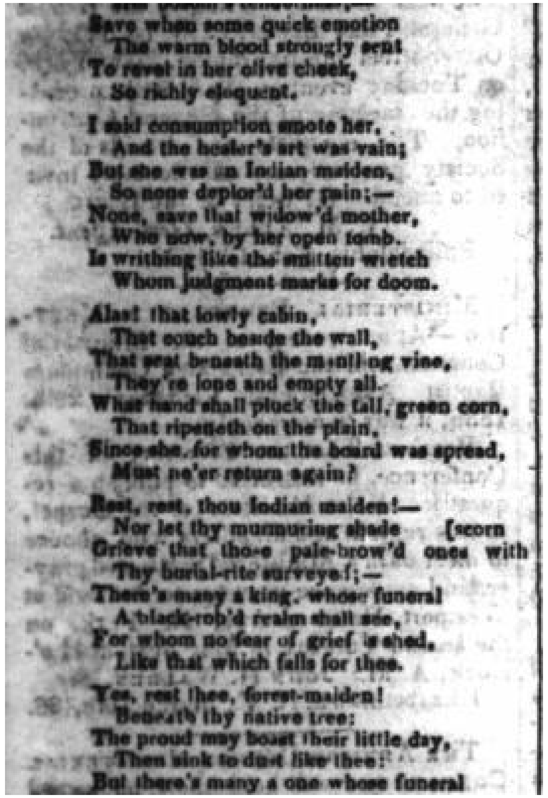
WASHINGTON HALL,
University of Virginia.

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 eulogy on his character, but the sorrowed countenances

good quality



Save when some quick emotion
The warm blood strongly sent
To revel in her olive cheek,
So richly eloquent.

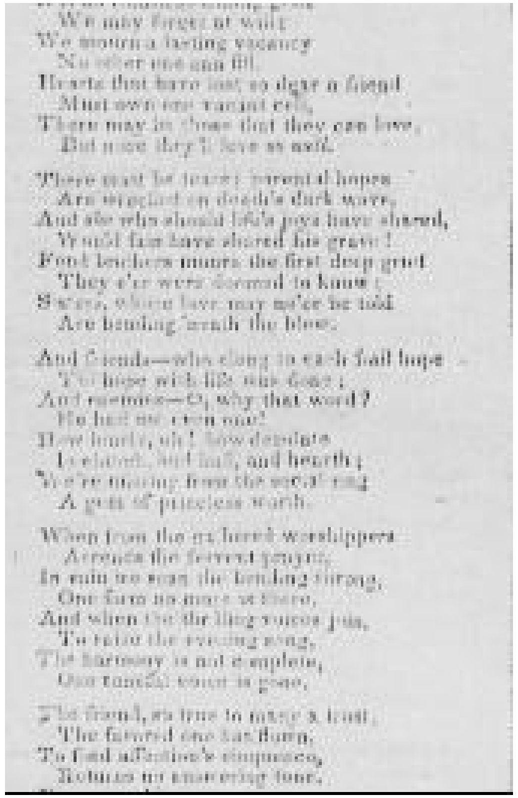
I said consumption smote her,
And the healer's art was vain;
But she was an Indian maiden,
So none deplor'd her pain;—
None, save that widow'd mother,
Who now, by her open tomb,
Is writhing like the smitten wretch
Whom judgment marks for doom.

Alas! that lowly cabin,
That couch beside the wall,
That seat beneath the mantling vine,
They're lone and empty all.
What hand shall pluck the tall, green corn,
That ripeneth on the plain,
Since she, for whom the board was spread,
Must ne'er return again?

Rest, rest, thou Indian maiden!—
Nor let thy murmuring shade
Grieve that those pale-brow'd ones with [scorn
Thy burial-rite surveyed;—
There's many a king, whose funeral
A black-rob'd realm shall see,
For whom no tear of grief is shed,
Like that which falls for thee.

Yes, rest thee, forest-maiden!
Beneath thy native tree:
The proud may boast their little day,
Then sink to dust like thee!
But there's many a one whose funeral

bleed-through



We may forget at will;
We mourn a lasting vacancy
No other one can fill.
Hearts that have lost so dear a friend
Must own one vacant cell,
There may be those that they can love,
But none they'll love so well.

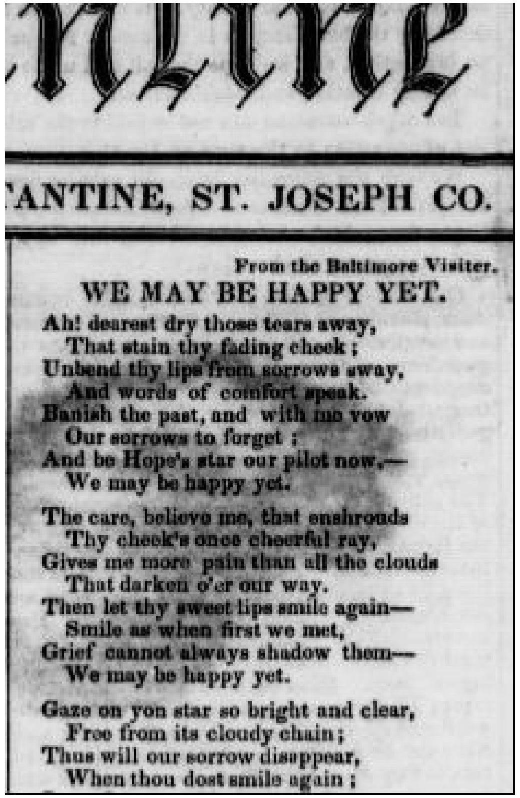
There must be [illegible] parental hopes
Are wrecked on death's dark wave,
And she who should life's joys have shared,
Would fain have shared his grave!
Fond brothers mourn the first deep grief
They e'er were doomed to know;
Sisters, whose love may ne'er be told
Are bending 'neath the blow.

And friends—who clung to each frail hope
Till hope with life was done;
And enemies—O, why that word?
He had not even one!
How lonely, oh! how desolate
Is church, and hall, and hearth;
We're missing from the social ring
A gem of priceless worth.

When from the gathered worshippers
Ascends the fervent prayer,
In vain we scan the bending throng,
One form no more is there,
And when the thrilling voices join,
To raise the evening song,
The harmony is not complete,
One tuneful voice is gone,

The friend, so true to many a trust,
The favored one has [illegible],
To find affection's eloquence,
Returns no answering tone.

low contrast



[illegible]NTINE, 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 away,
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 cheek'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 yon star so bright and clear,
Free from its cloudy chain;
Thus will our sorrow disappear,
When thou dost smile again;

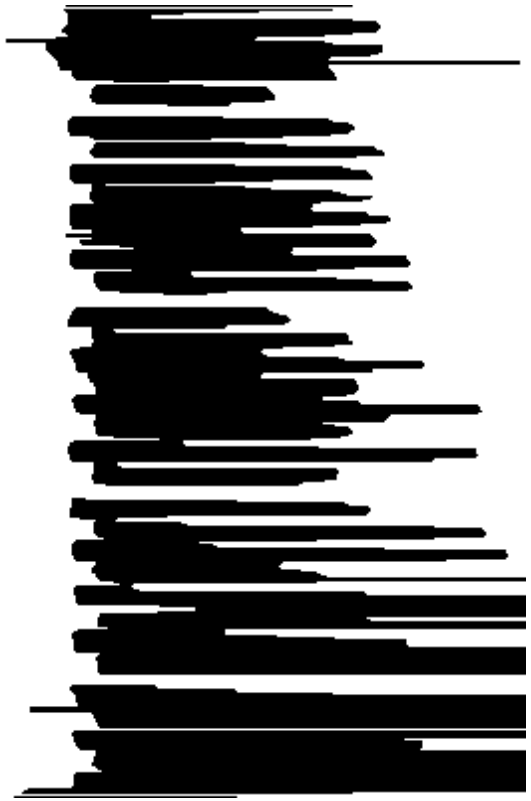
occluding "blobs"

IMPROVED APPROACH

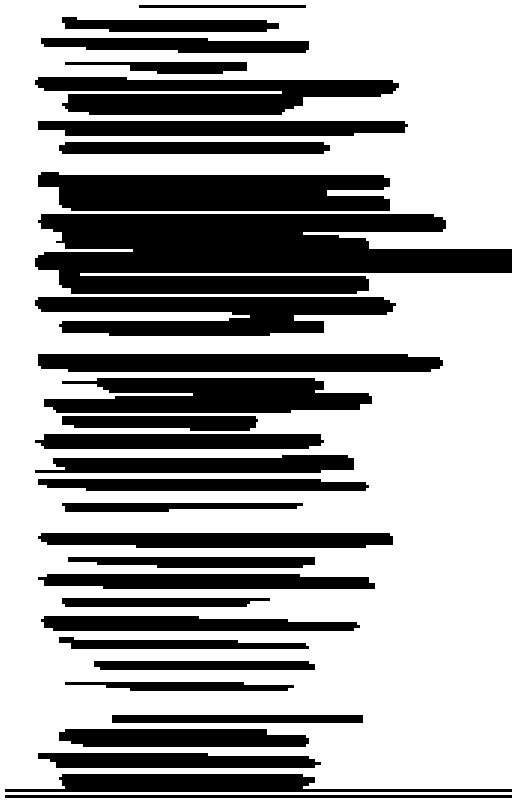
Improve our binarization of the image into "object" and "background" information (represented as black and white pixels, respectively) and extract visual structures with a noise-tolerant text line segmentation approach

Results from Initial Approach

good quality



bleed-through

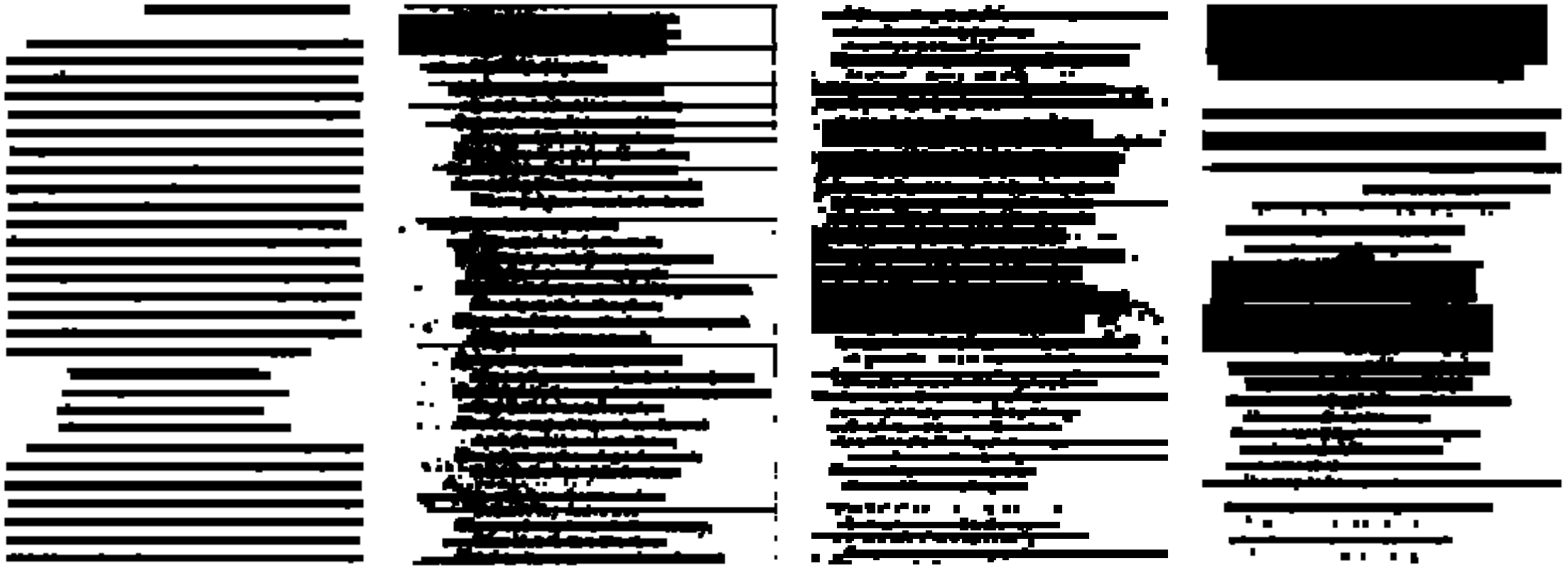


low contrast

occluding “blobs”

Aida

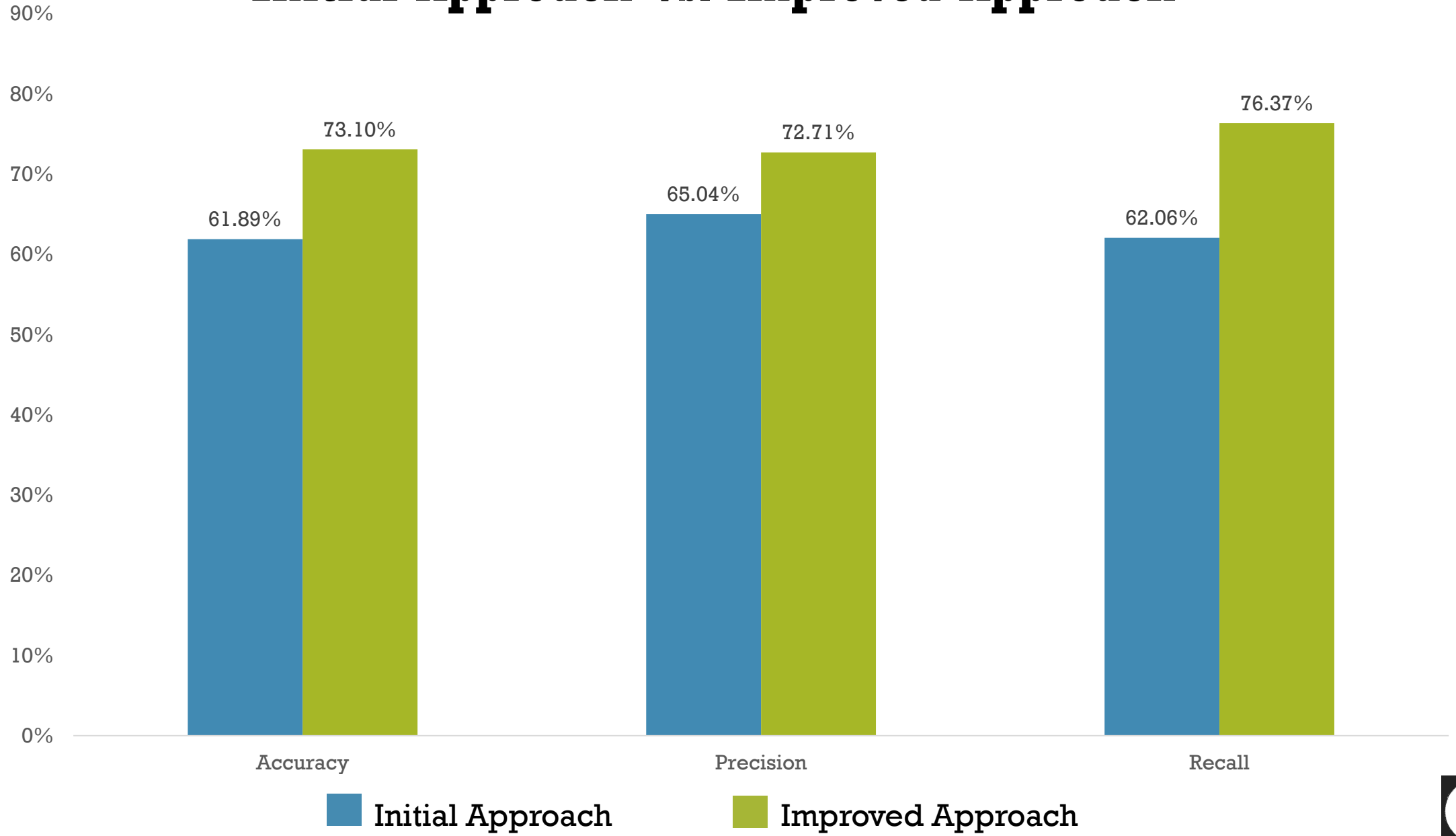
Results from Improved Approach



good quality | bleed-through | low contrast | occluding "blobs"

Aida

Initial Approach vs. Improved Approach



Aida

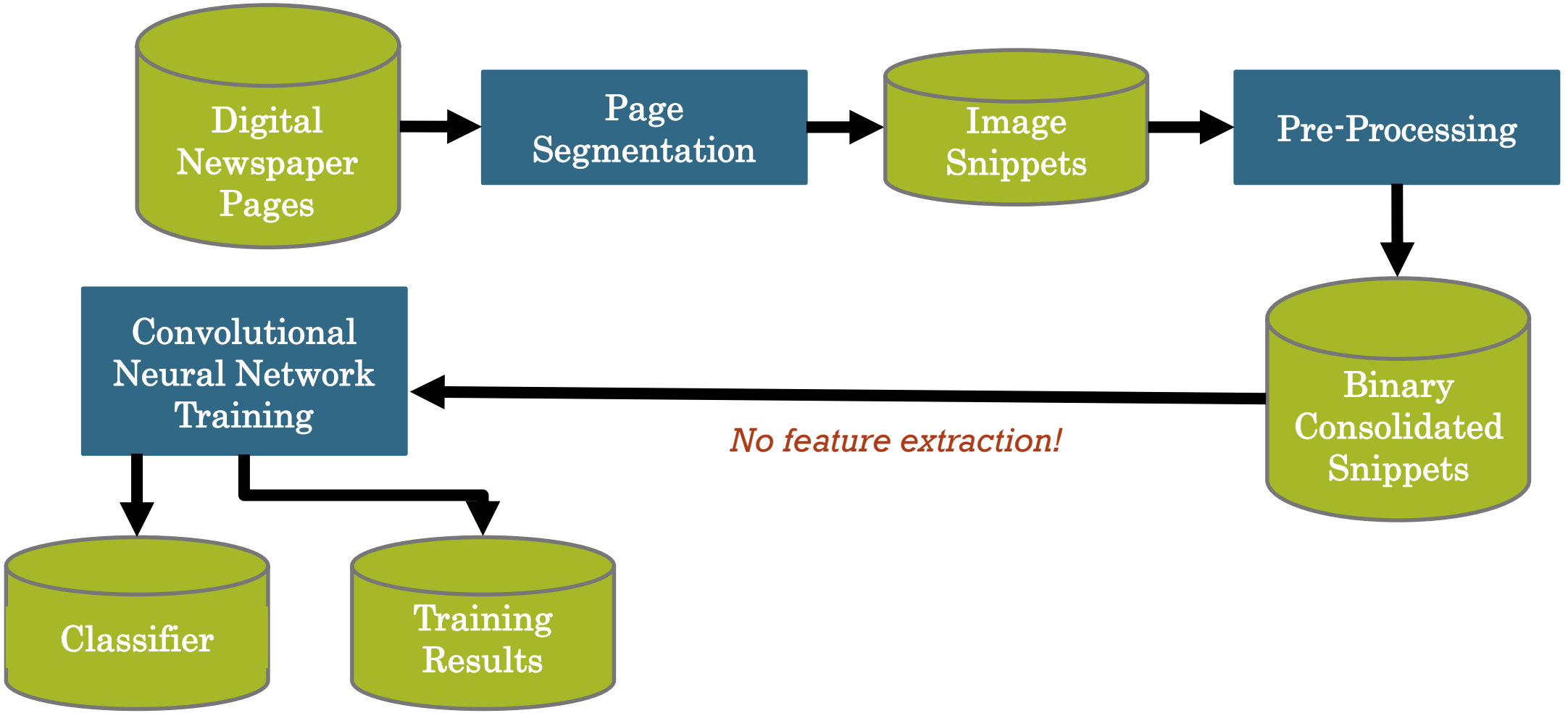
ALTERNATIVE APPROACH

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

Initial results of accuracy achieved 92.4%

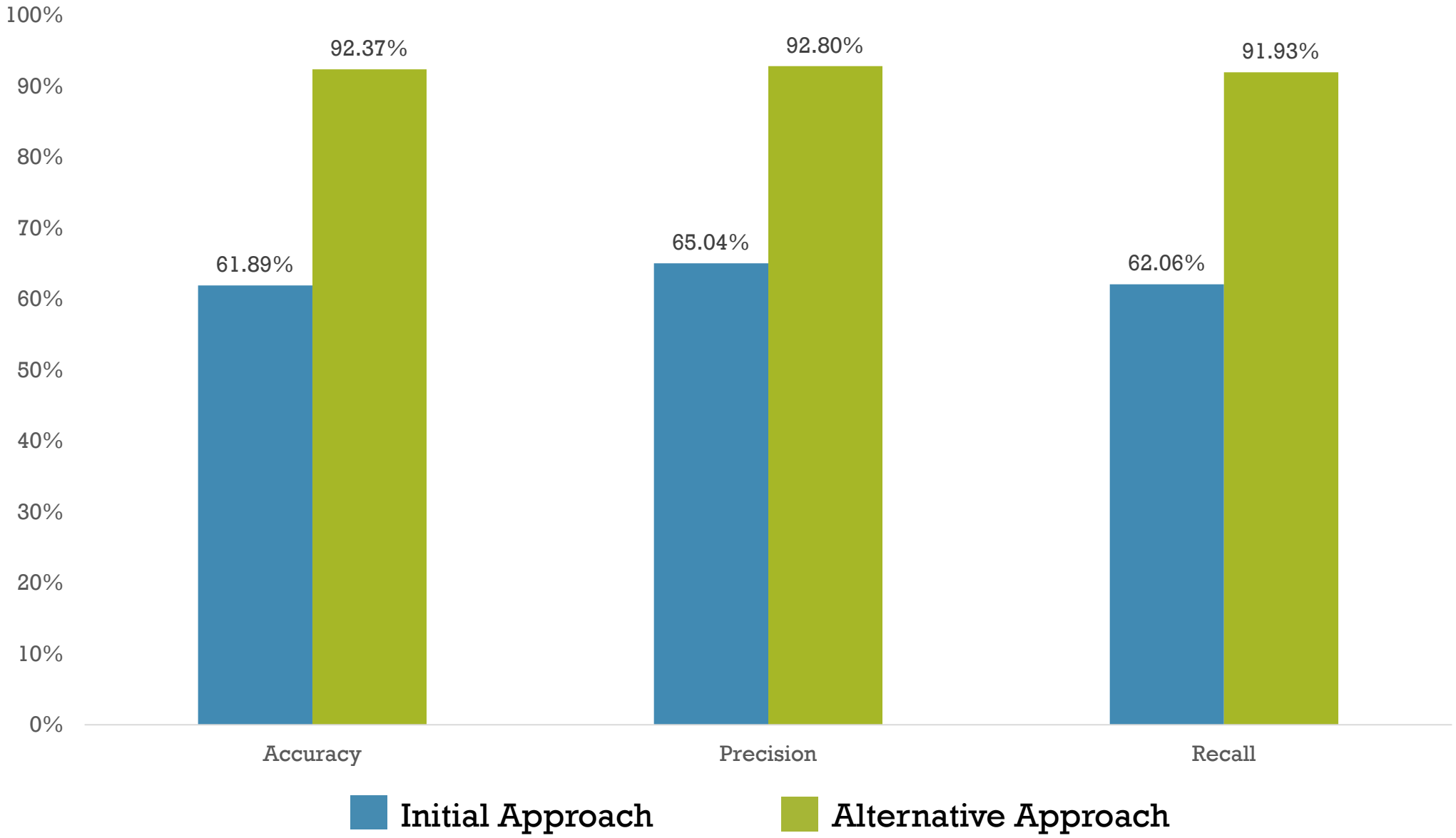
- precision = 92.8%; recall = 91.9%

Aida


Digital Newspaper Pages
Page Segmentation
Image Snippets
Pre-Processing
Binary Consolidated Snippets
Convolutional Neural Network Training
No feature extraction!
Classifier
Training Results

Aida

Initial Approach vs. Alternative Approach



Aida

LAYOUT ANALYSIS

LAYOUT ANALYSIS

Layout analysis and zoning performed by OCR systems as a pre-processing step for text recognition

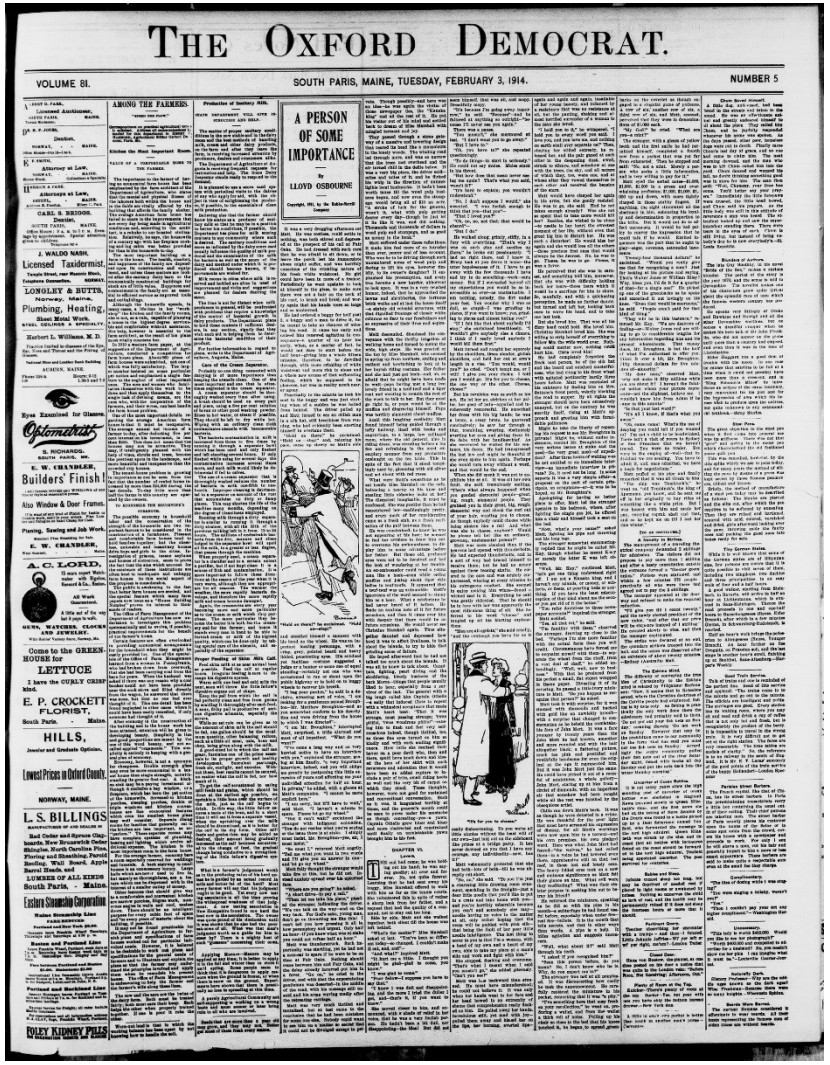
- E.g., built in to OCR web services such as ABBYY Cloud and in the open source Tesseract OCR engine

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

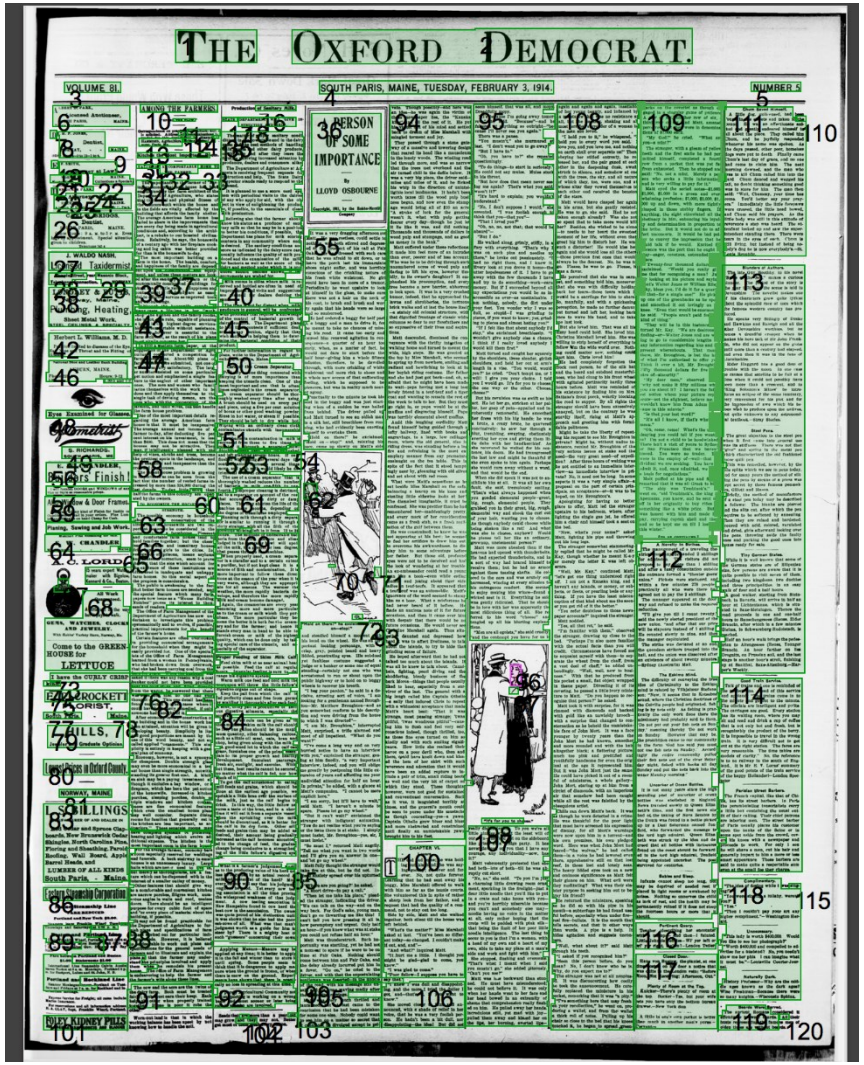
- Zoning coordinates for strings, text lines, some larger text blocks, and page-level information

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

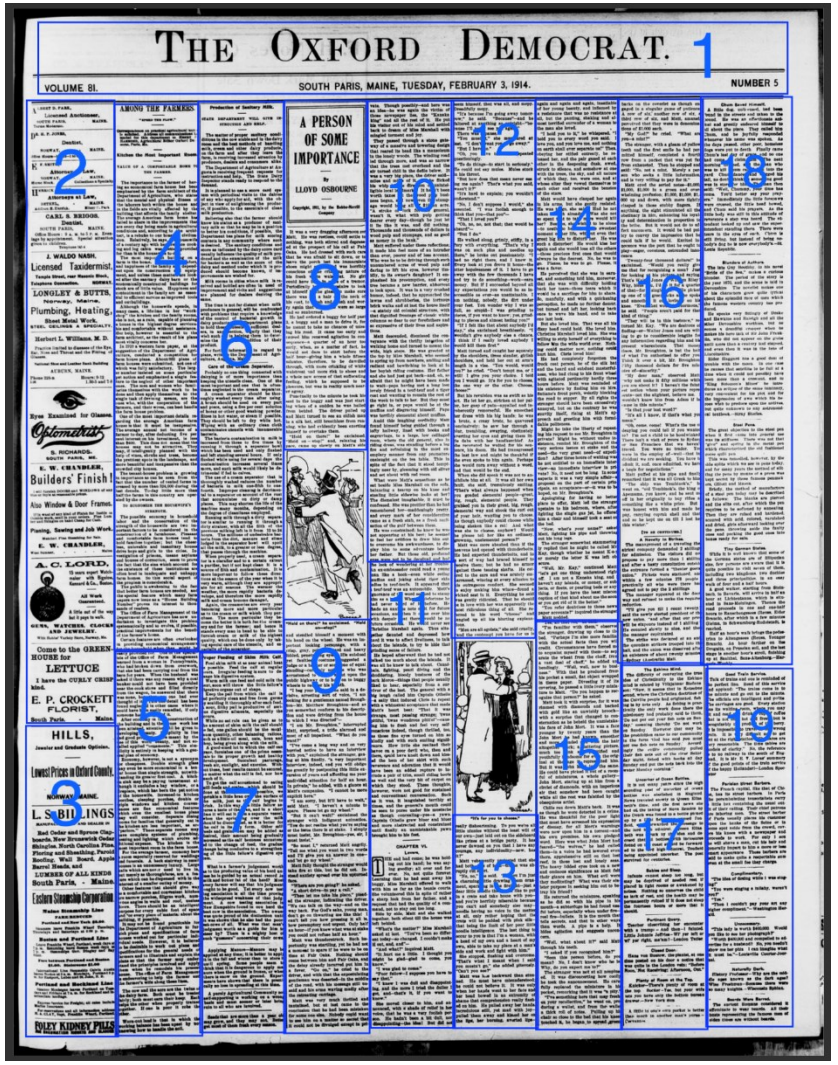
Aida



newspaper page from Chronicling America

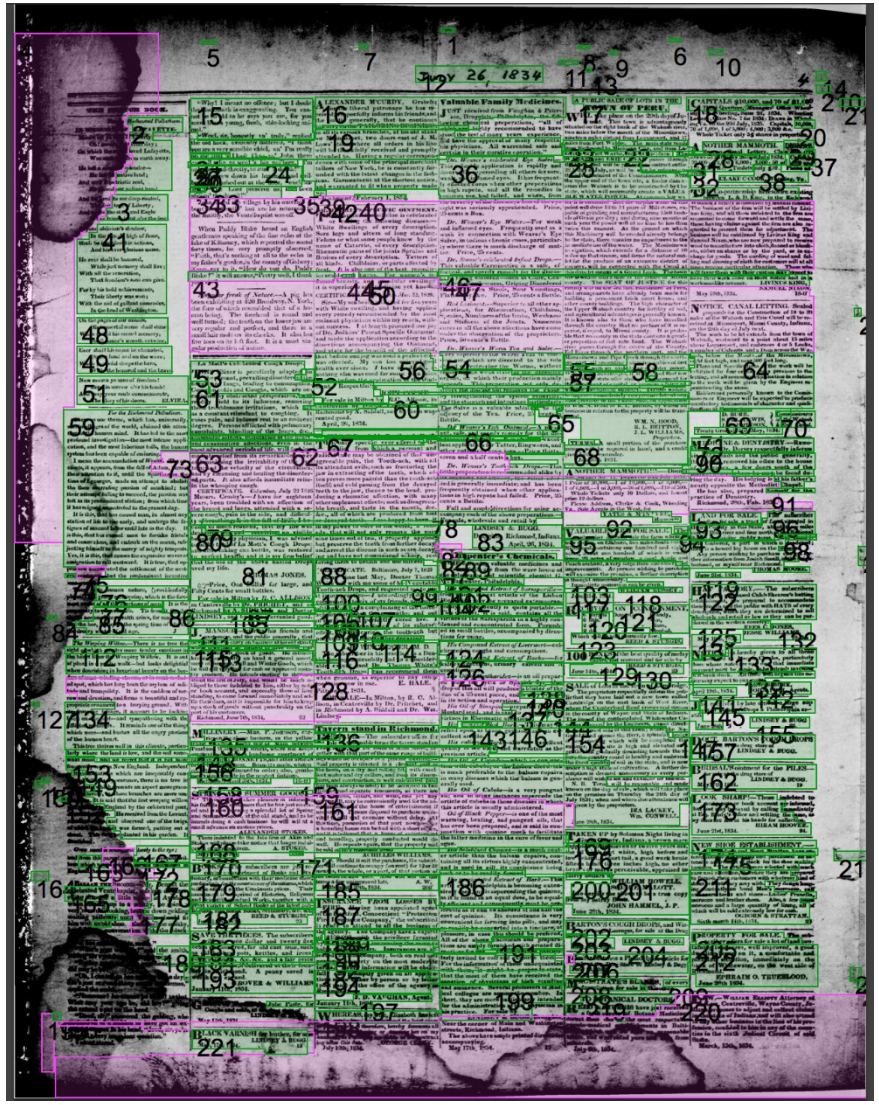


zones identified by Abbyy FineReader

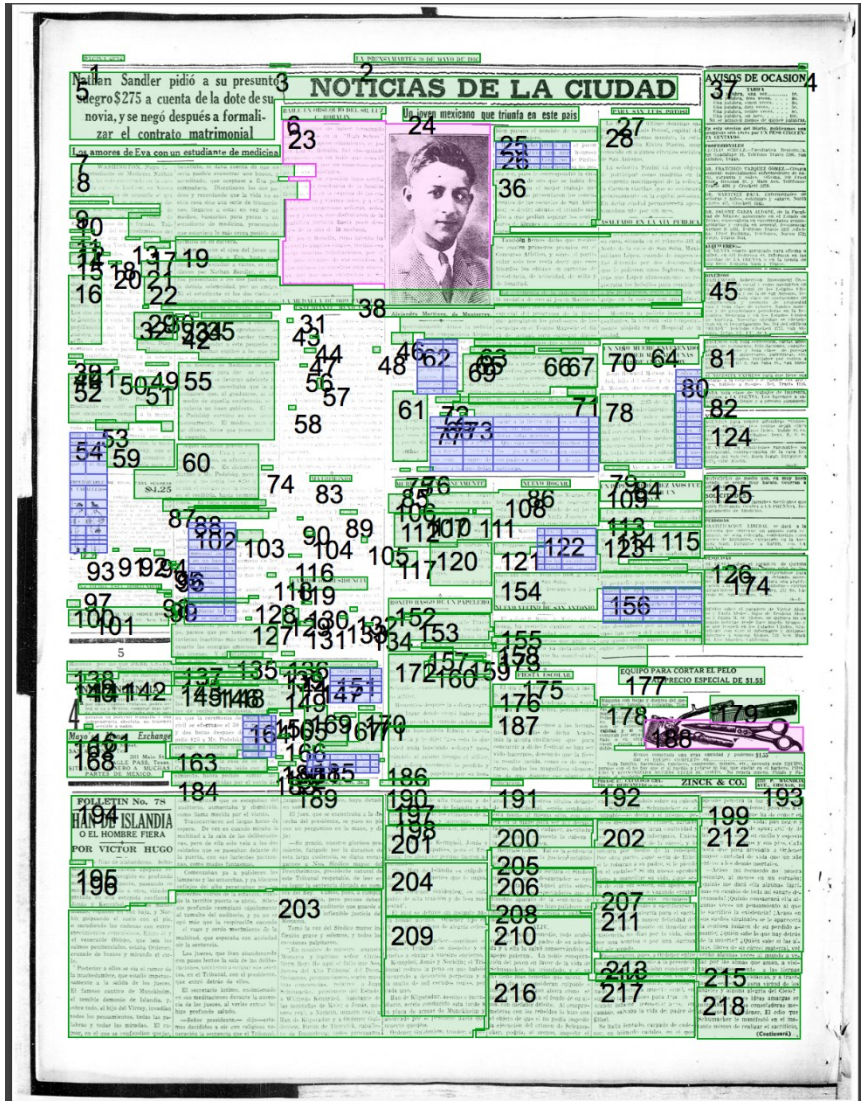


text blocks recorded in OCR XML on Chronicling America

Aida



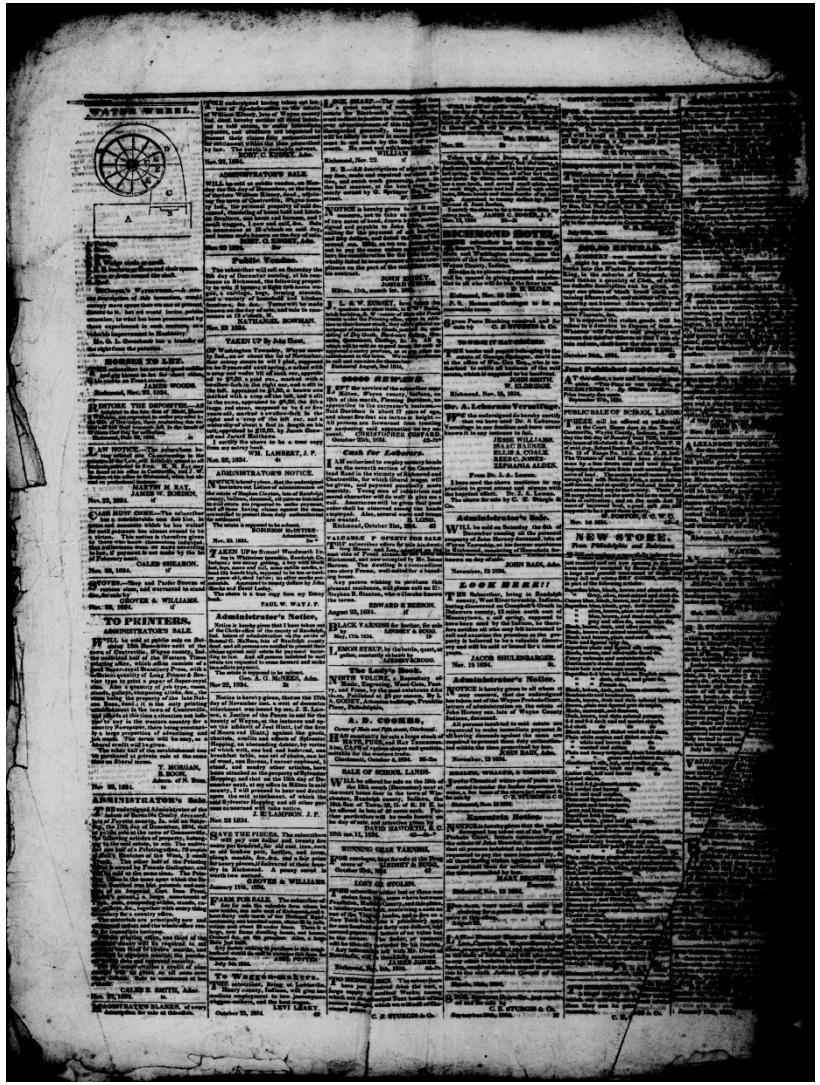
challenges to zoning:
damage & noise



challenges to zoning: contrast
and range effect

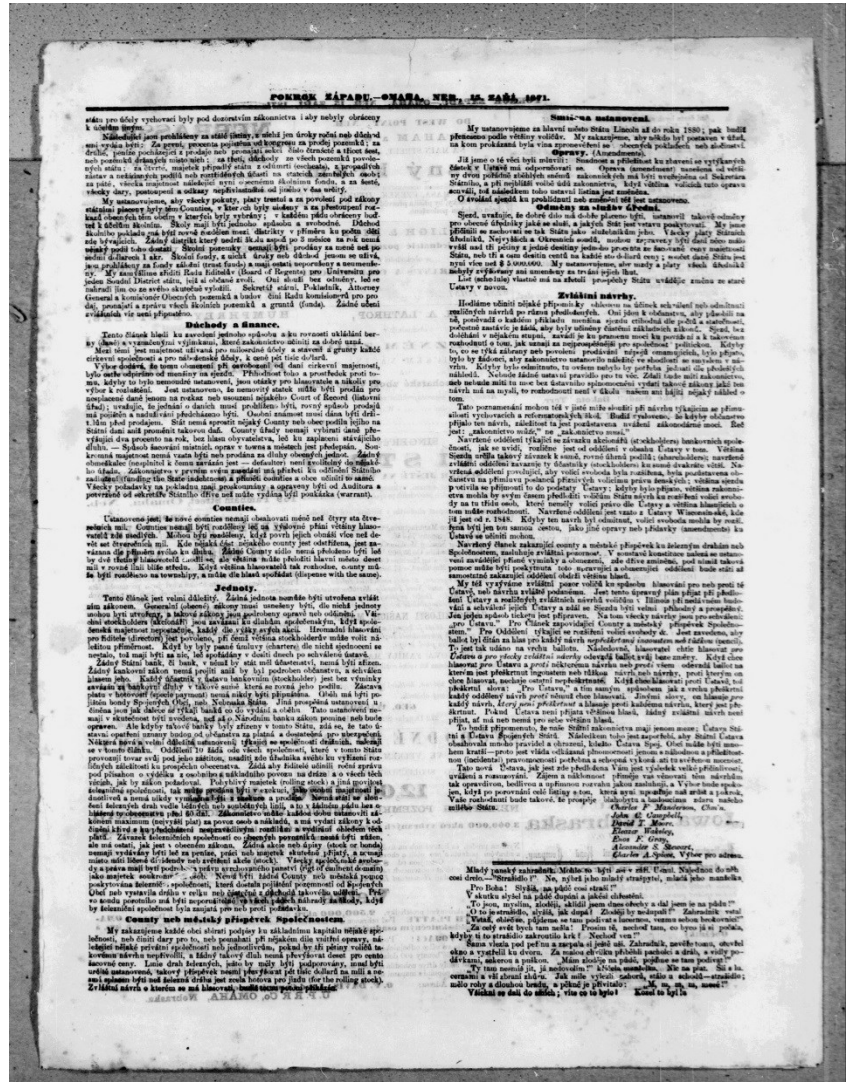
Aida

CORPUS-WIDE ANALYSIS

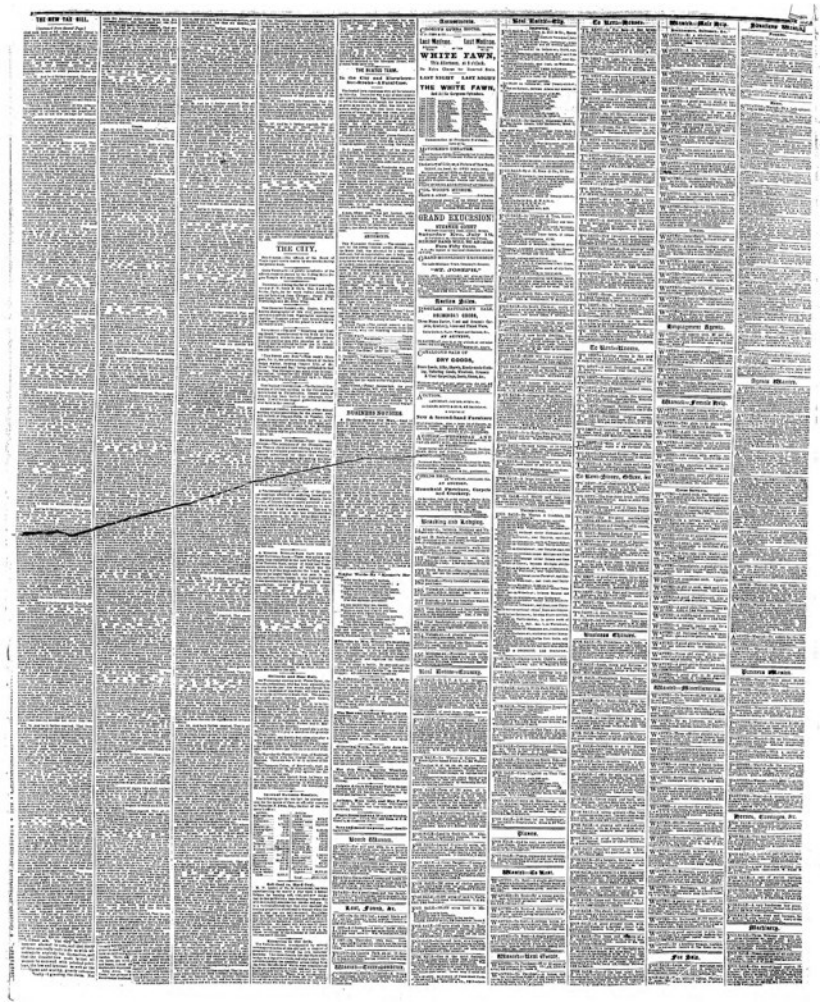


sn82016416_1834-11-29_ed-1_seq-4

sn90059649_1916-08-04_ed-1_seq-8

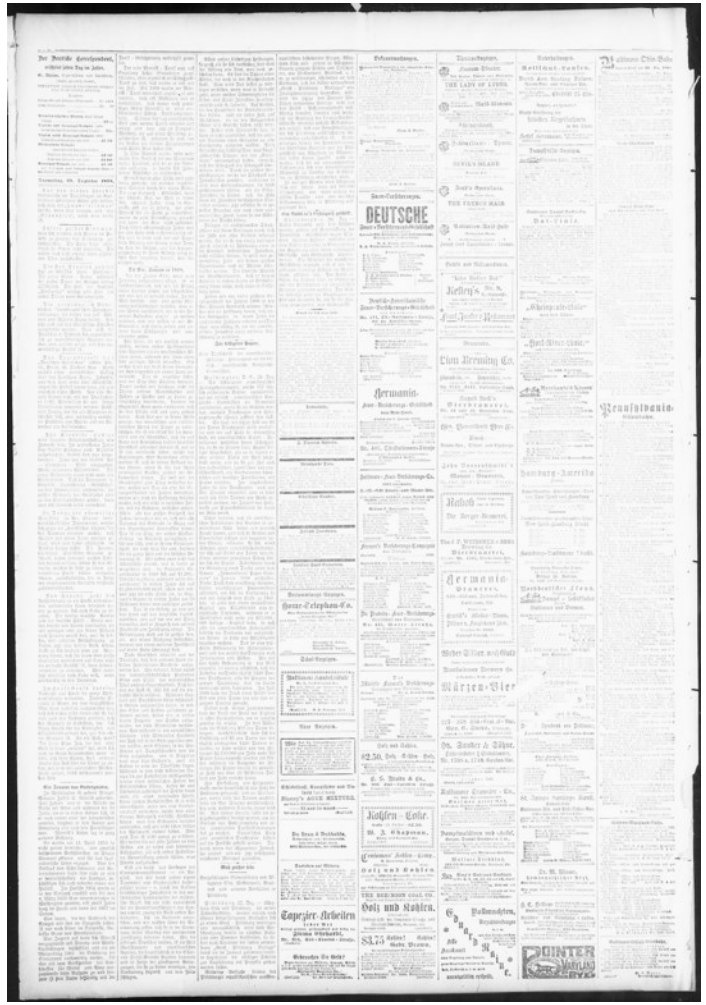


sn83045348_1871-09-15_ed-1_seq-3



Newspaper page showing high contrast

Newspaper page showing average contrast



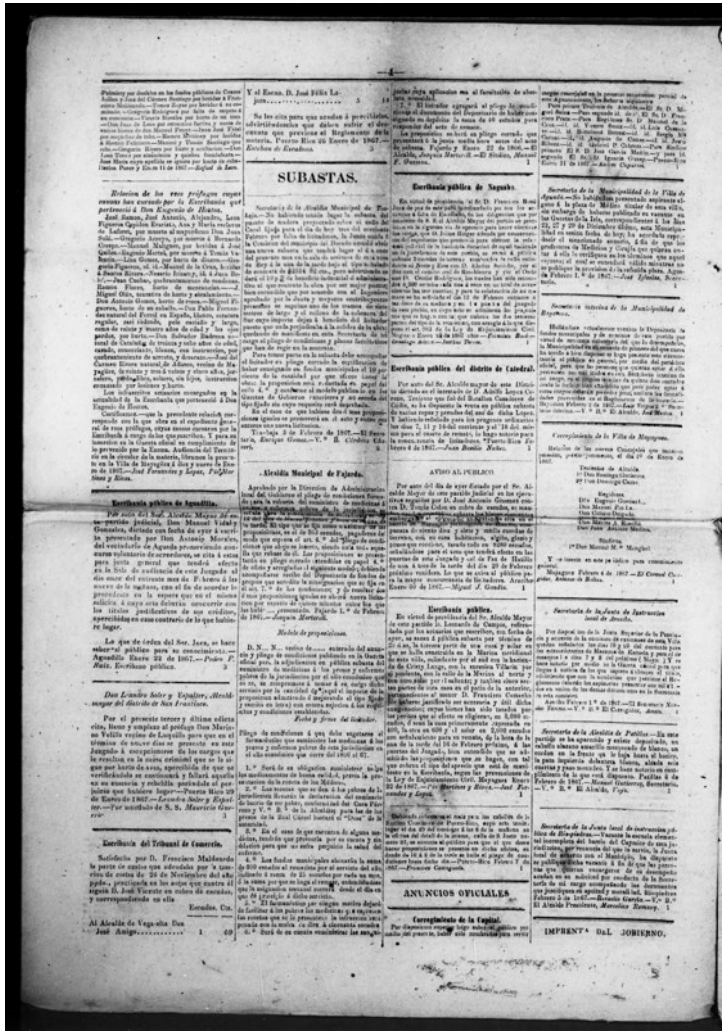
Newspaper page showing low contrast

Aida

THE BRYAN DAILY EAGLE.

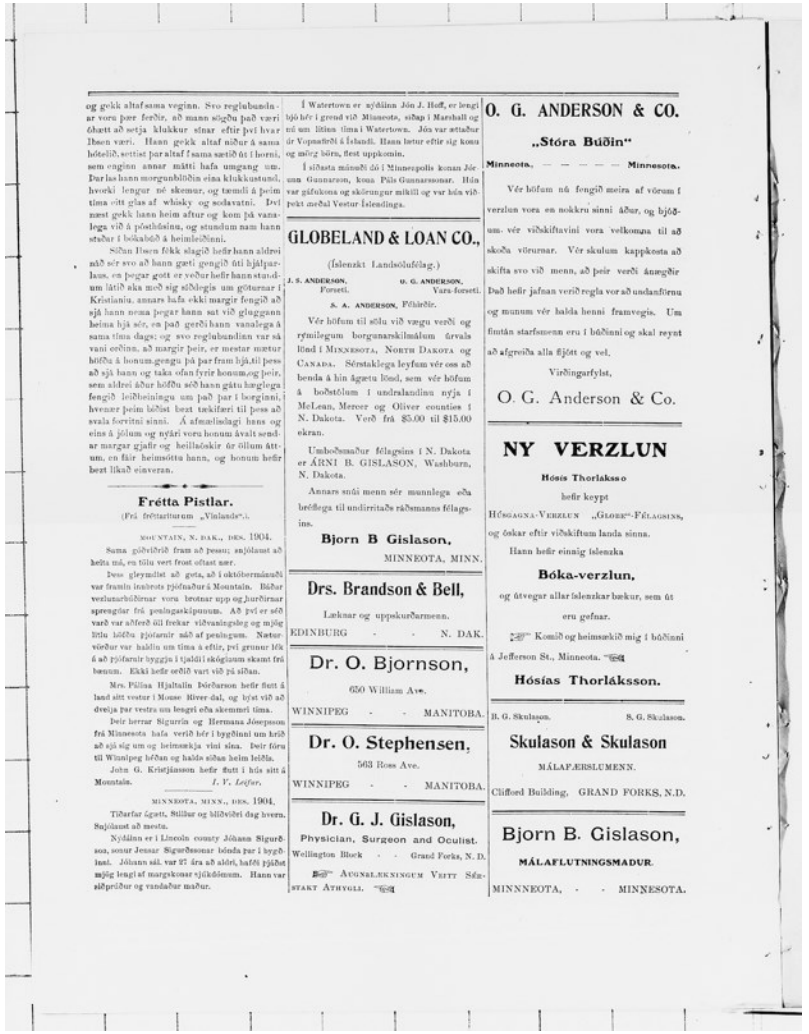
BRYAN, TEXAS, SATURDAY MORNING, APRIL 18, 1896.

Newspaper page showing high range-effect



SUBASTAS.

Newspaper page showing average range-effect



Frétta Pistlar.

O. G. ANDERSON & CO.

„Stóra Búðin"

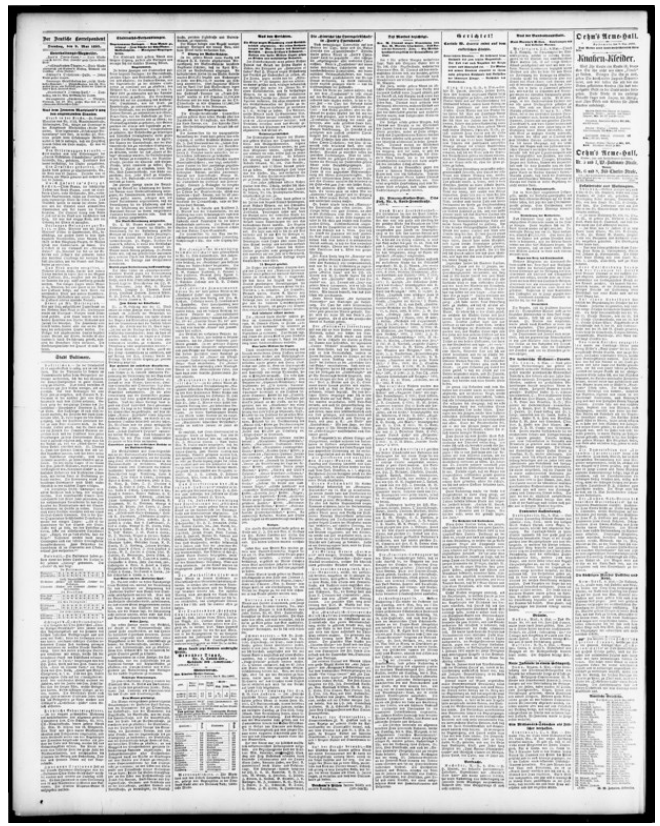
NY VERZLUN

Newspaper page showing no range-effect

Aida

Newspaper page showing severe orientation skew

Newspaper page showing some orientation skew



Newspaper page showing no orientation skew

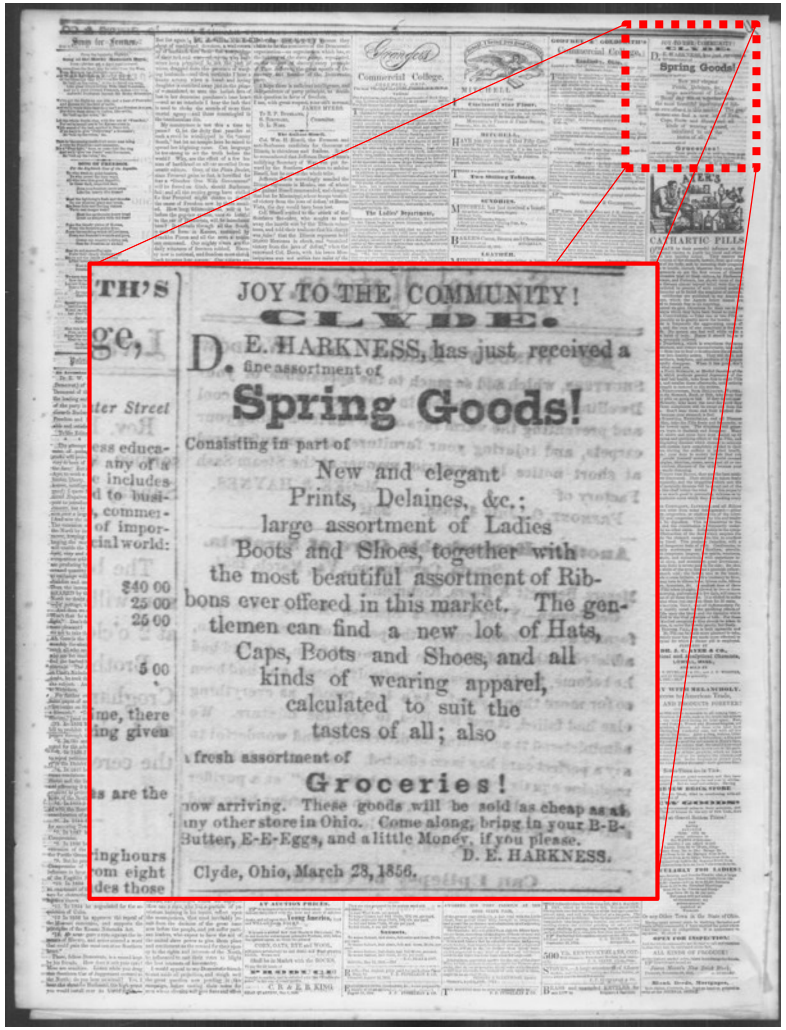
Aida

AVISOS.

CATECISMO DEL Sistema-métrico decimal

LIBROS.

Newspaper page showing severe bleed-through



JOY TO THE COMMUNITY!

CLYDE.

D. E. HARKNESS, has just received a fine assortment of

Spring Goods!

Consisting in part of

New and elegant Prints, Delaines, &c.; large assortment of Ladies Boots and Shoes, together with the most beautiful assortment of Ribbons ever offered in this market. The gentlemen can find a new lot of Hats, Caps, Boots and Shoes, and all kinds of wearing apparel, calculated to suit the tastes of all; also a fresh assortment of

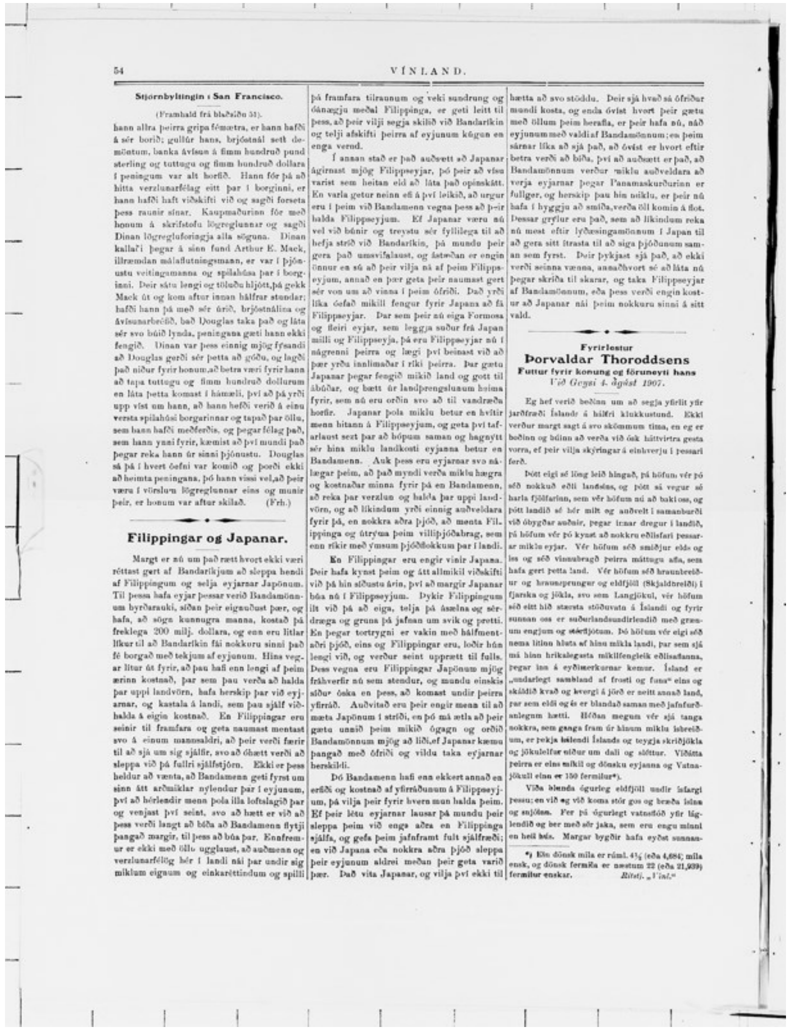
Groceries!

now arriving. These goods will be sold as cheap as at any other store in Ohio. Come along, bring in your B-B-Butter, E-E-Eggs, and a little Money, if you please.

D. E. HARKNESS.

Clyde, Ohio, March 23, 1856.

Newspaper page showing some bleed-through



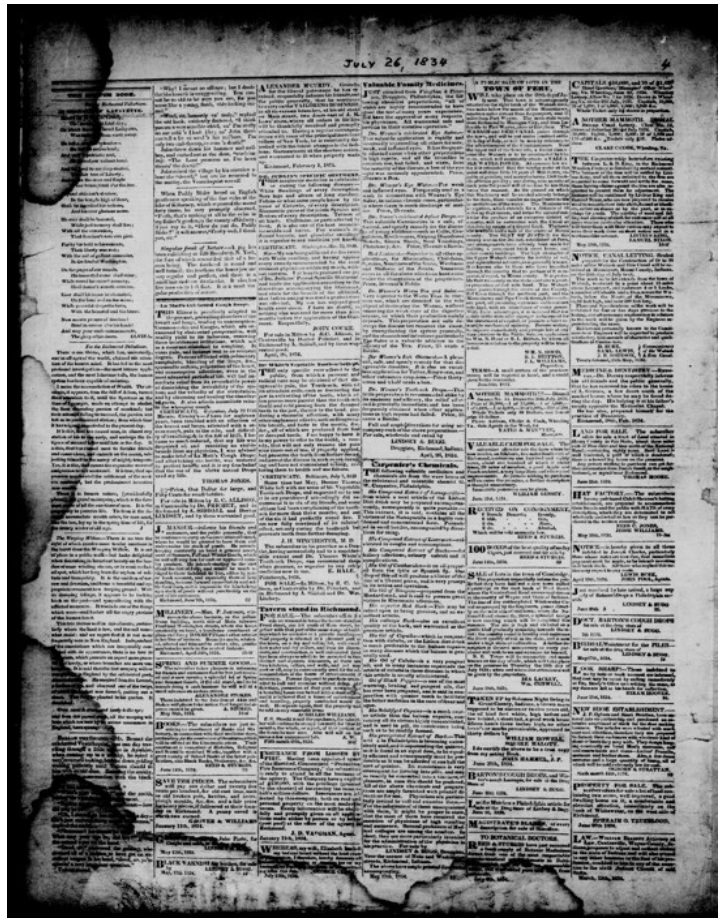
VÍNLAND.

Stjórnbyltingin í San Francisco.

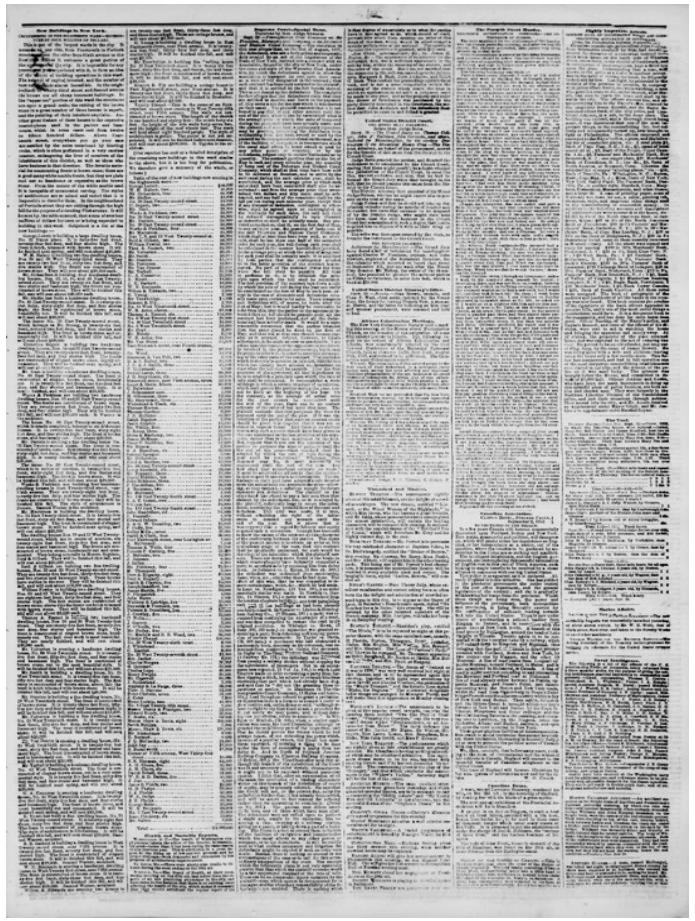
Filippingar og Japanar.

Newspaper page showing no bleed-through

Aida



Newspaper page showing severe noisiness



Newspaper page showing some noisiness

Newspaper page showing no noisiness

Aida

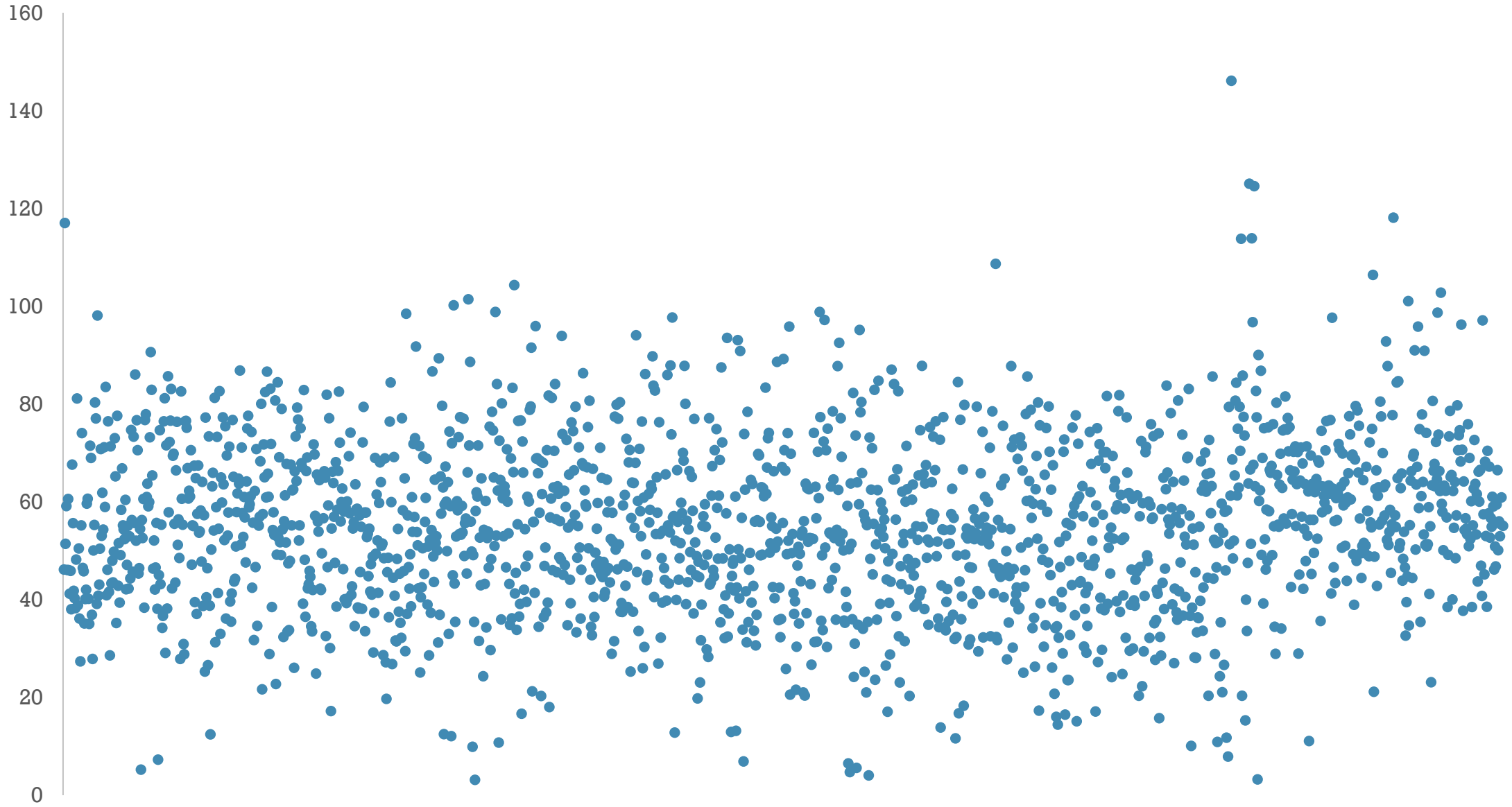
EVALUATION FEATURES

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

- Contrast
- Range effect
- Orientation skew
- Noise

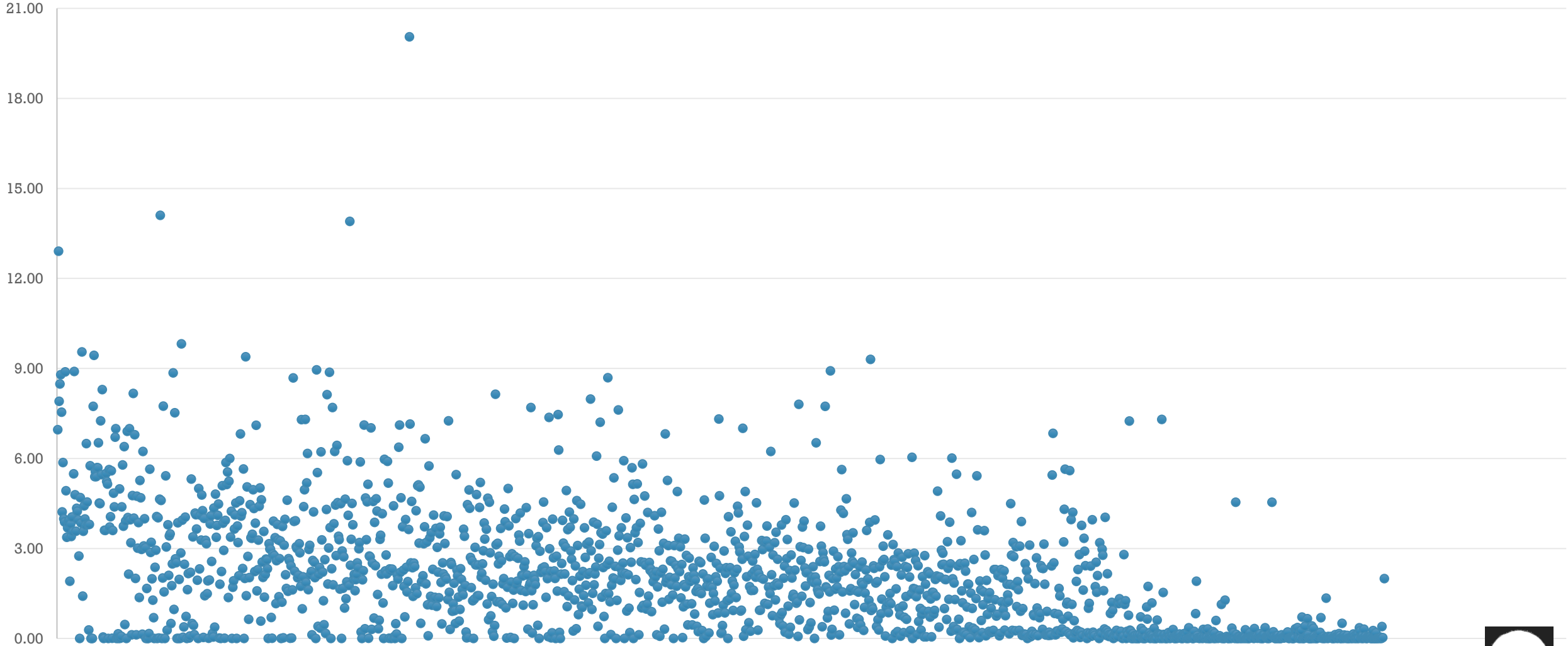
Aida

Contrast
Test Set



Aida

Range Effect
Test Set



Aida

ANALYSIS: CONTRAST, RANGE EFFECT

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

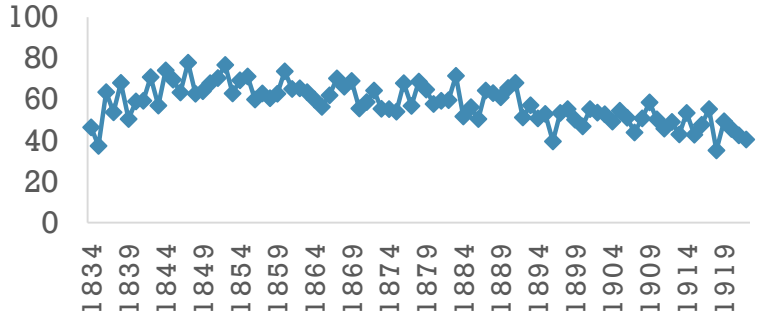
- The lower the contrast value, the worse the contrast
- ~ 6-7% images have bad contrast

Range effect becomes challenging when > 3

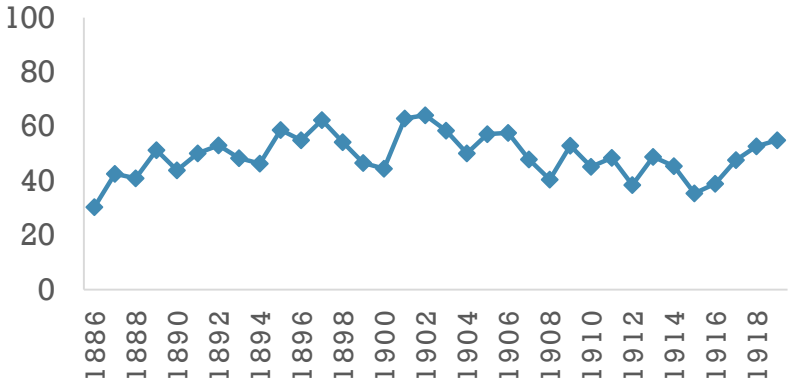
- The ideal value for range effect is 0 (= no range effect is present)
- ~ 25% images have challenging range effect

Aida

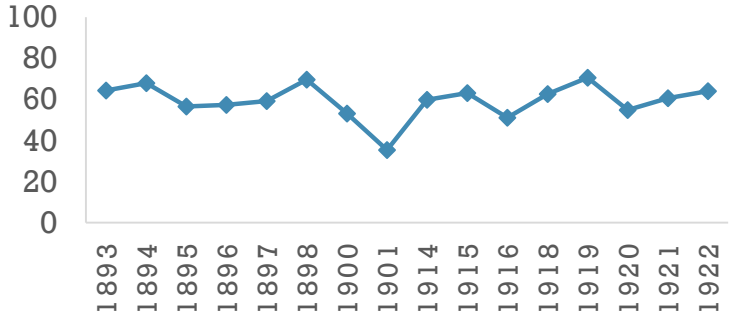
CONTRAST: ENGLISH



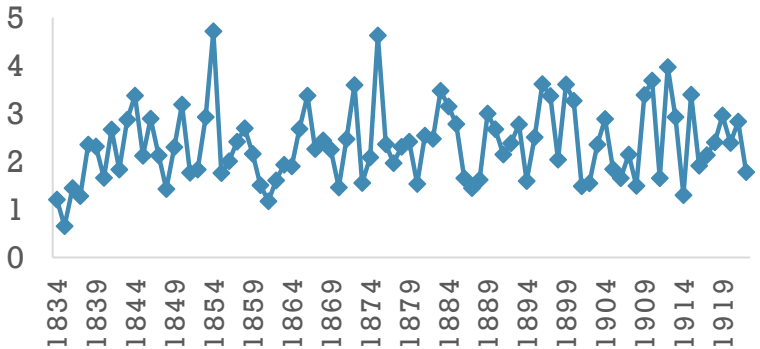
CONTRAST: POLISH



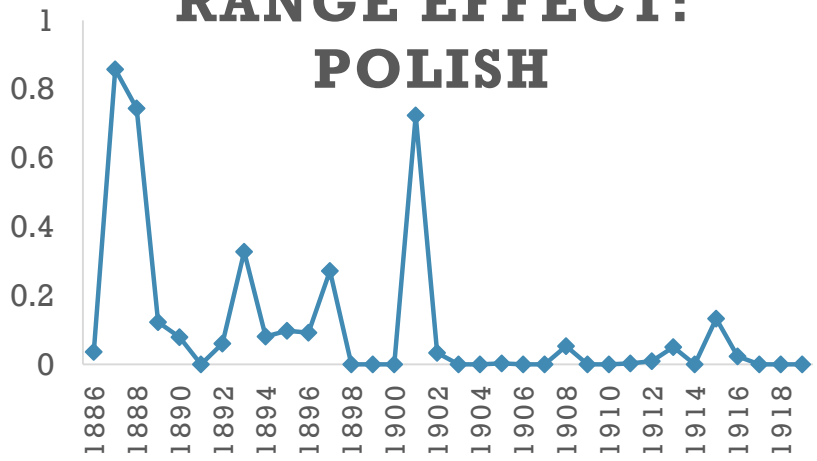
CONTRAST: NORWEGIAN



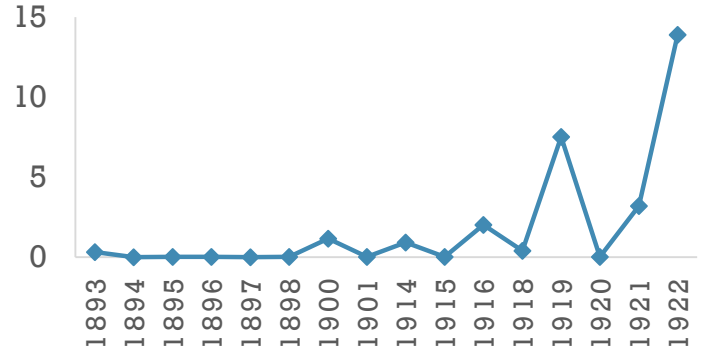
RANGE EFFECT: ENGLISH



RANGE EFFECT: POLISH



RANGE EFFECT: NORWEGIAN



Aida

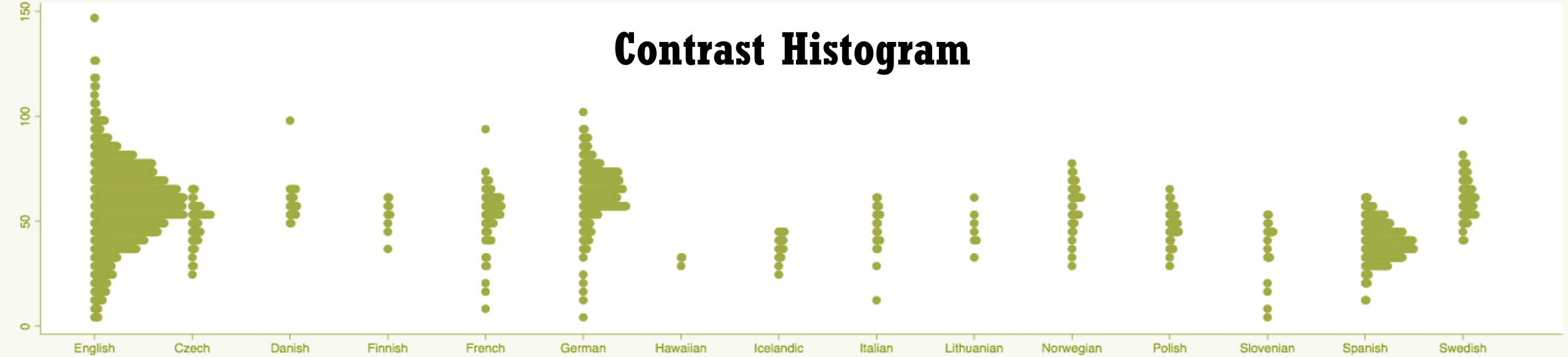
ANALYSIS 2: CONTRAST, RANGE EFFECT

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

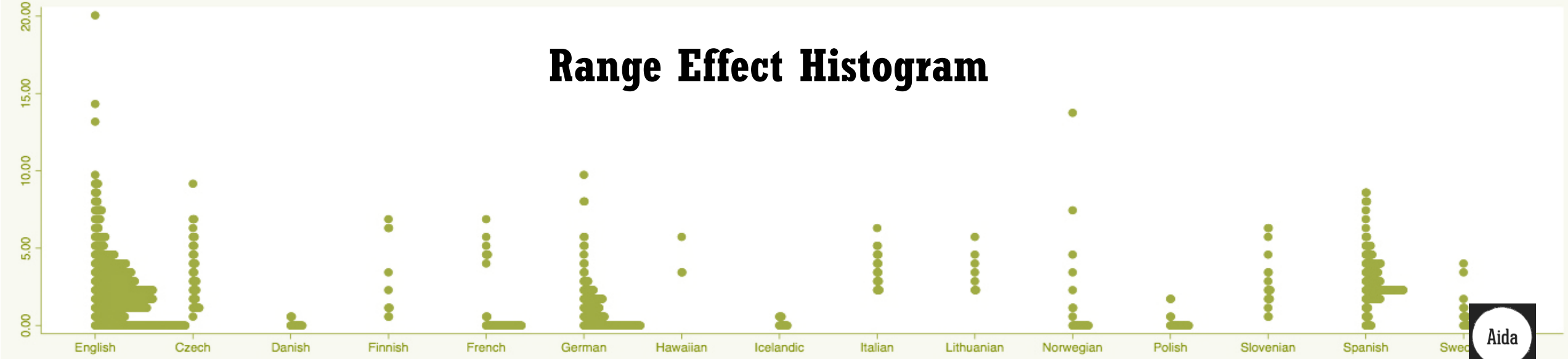
Range effect, on the other hand, not only varies across the different languages, it also changes over time for each language

Aida

Contrast Histogram



Range Effect Histogram



ANALYSIS 3: CONTRAST, RANGE EFFECT

From the histograms:

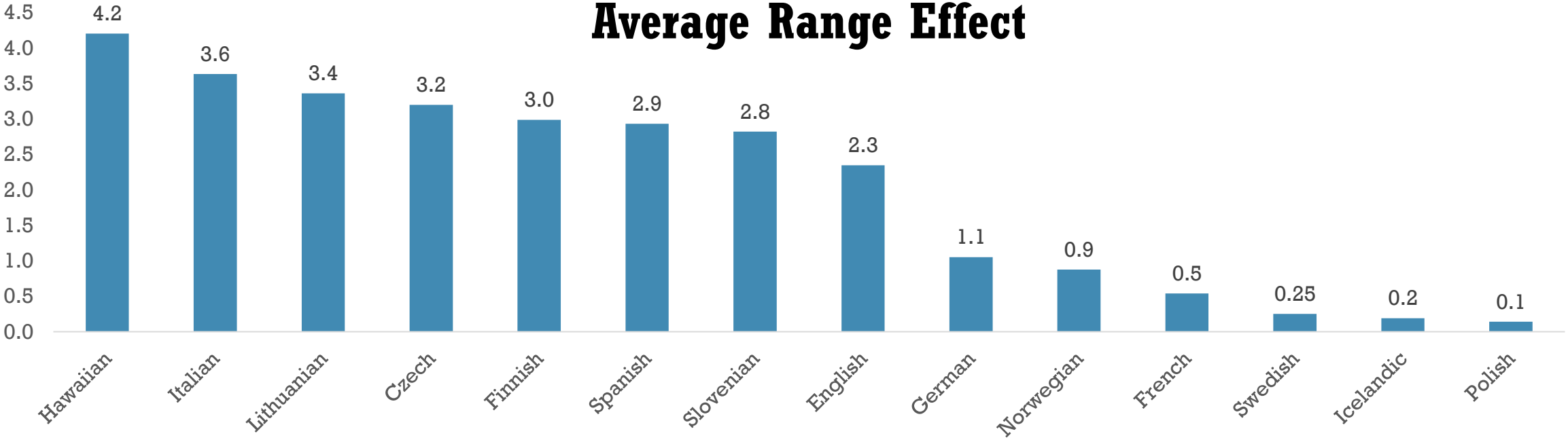
The data on Contrast are fairly symmetric, while the Range Effect data are right-skewed

A lot of outliers in our Range effect data

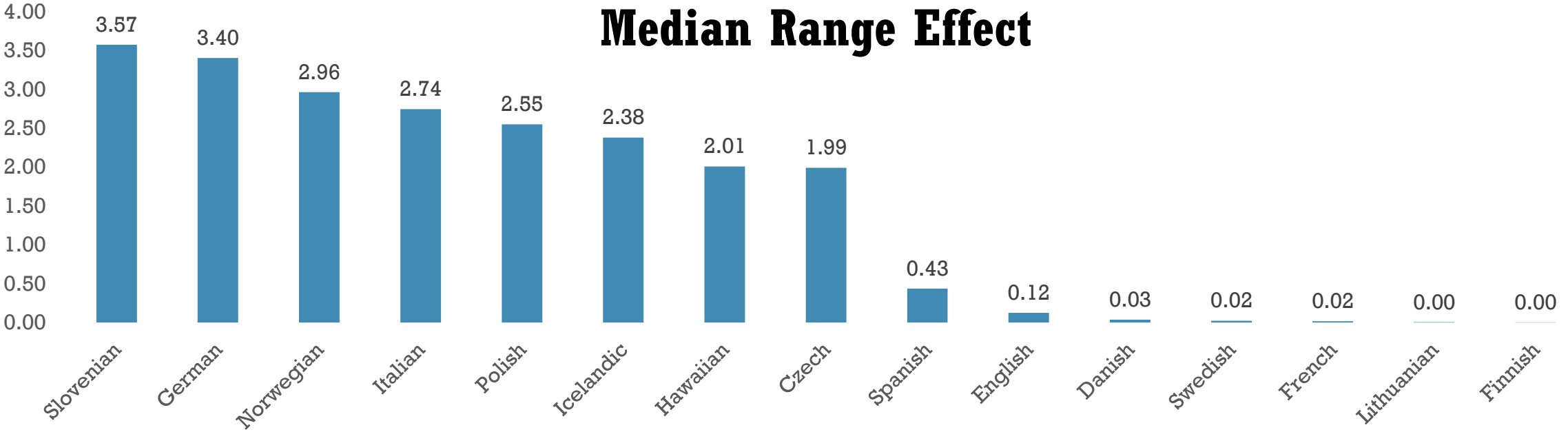
- The presence of outliers in our range effect data indicate that the mean (average) would not provide a good estimate: instead, median would give us a good estimate to evaluate the center of the data

Aida

Average Range Effect



Median Range Effect



Aida

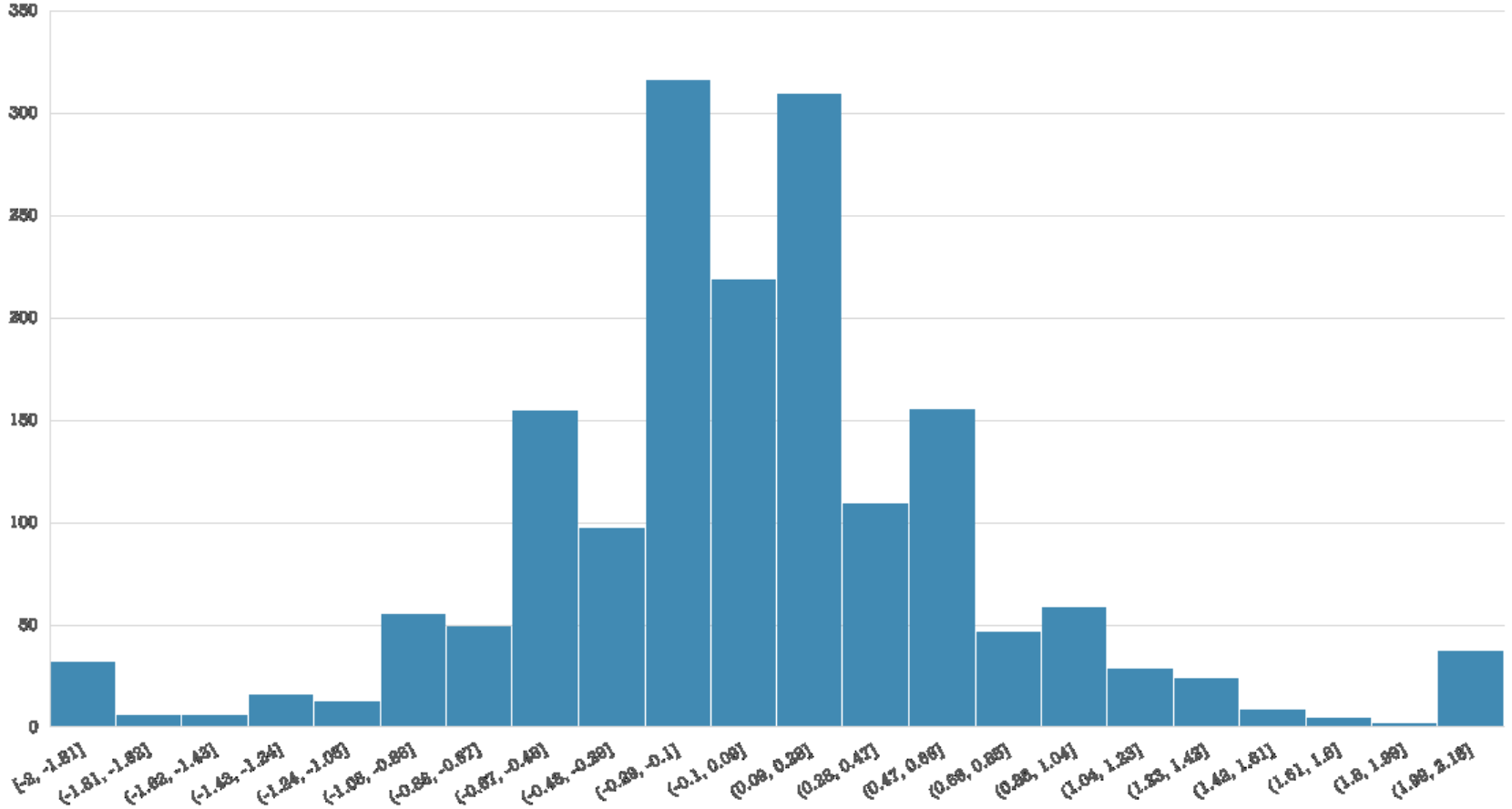
ANALYSIS: ORIENTATION SKEW

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.

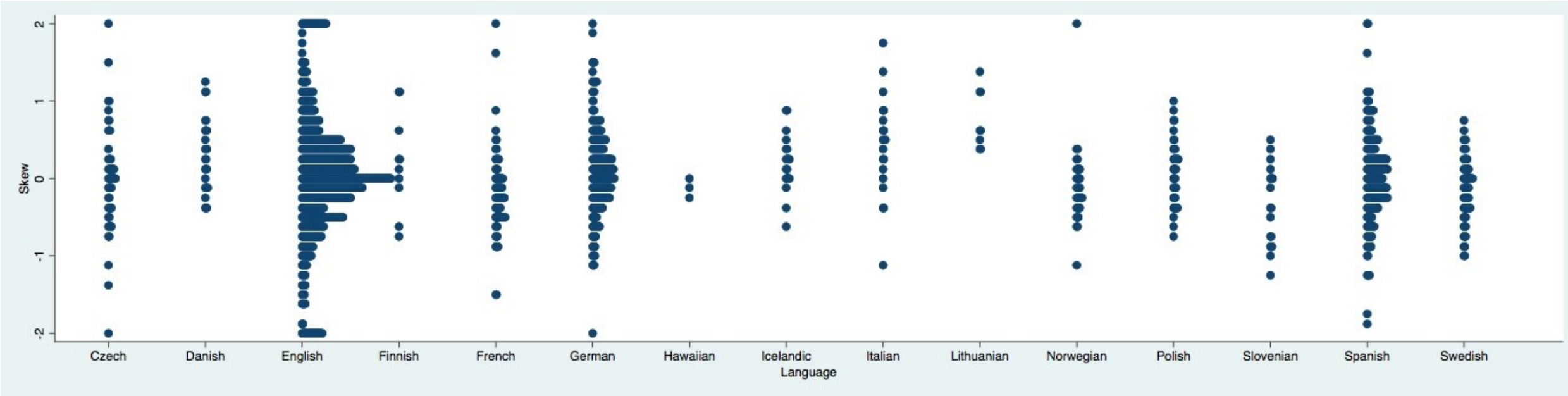
Aida

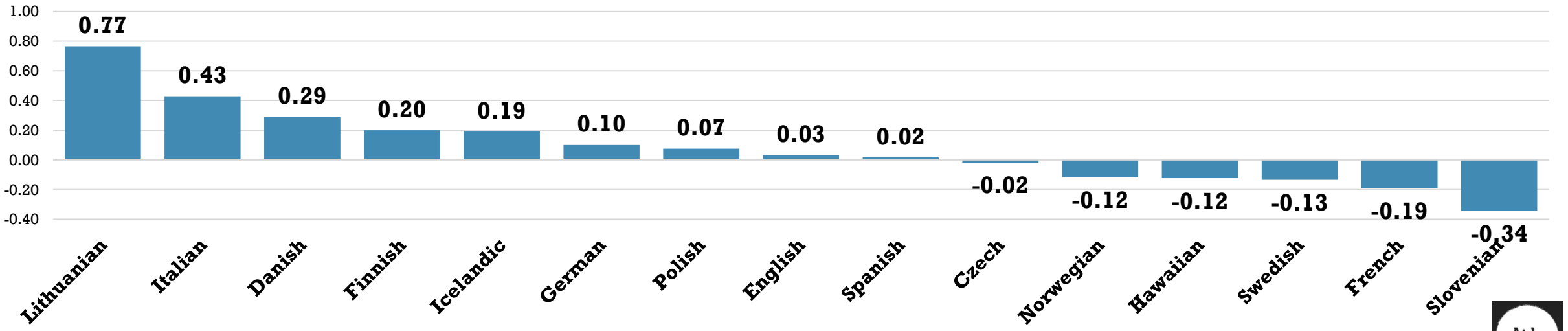
Distribution of Orientation Skew Test Set



Aida

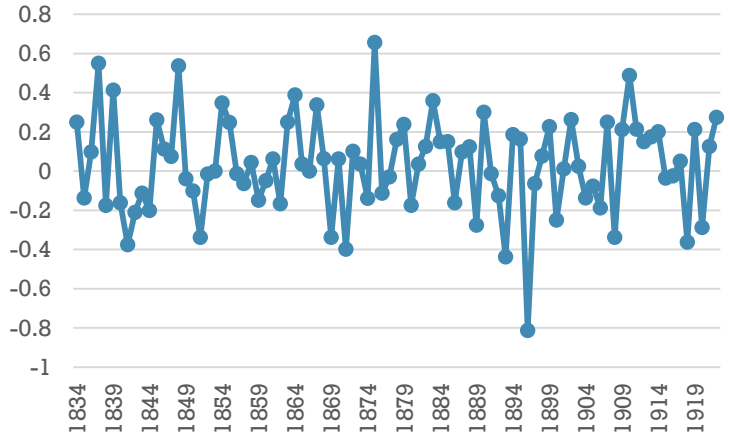
Views of Orientation Skew



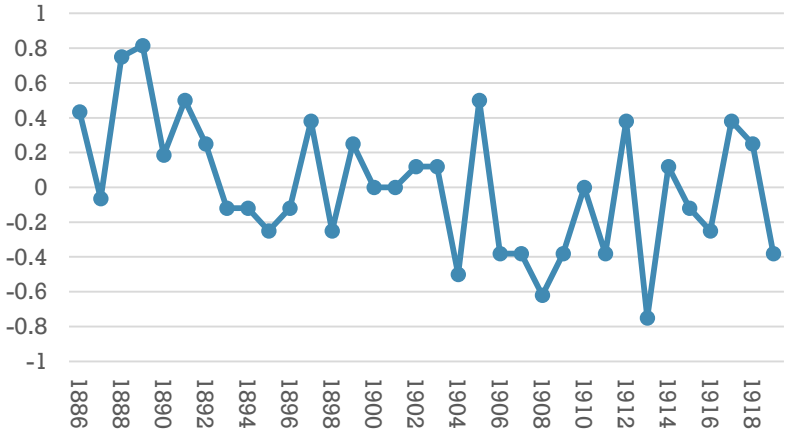


Aida

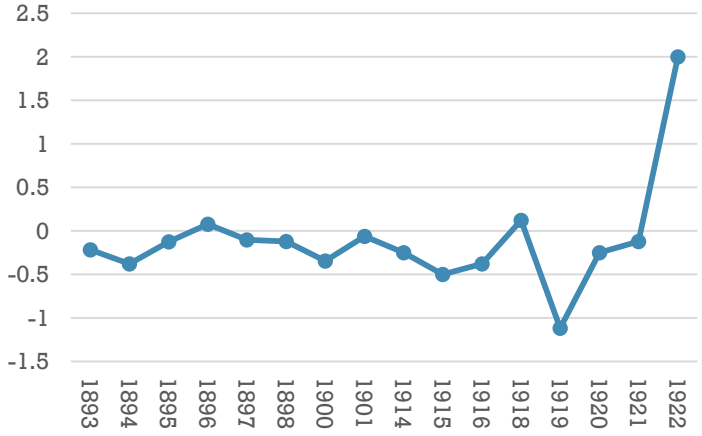
ORIENTATION SKEW: ENGLISH



ORIENTATION SKEW: POLISH



ORIENTATION SKEW: NORWEGIAN



Aida

Andrew Prescott
@Ajprescott

Follow

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

728 Retweets **1,467** Likes

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

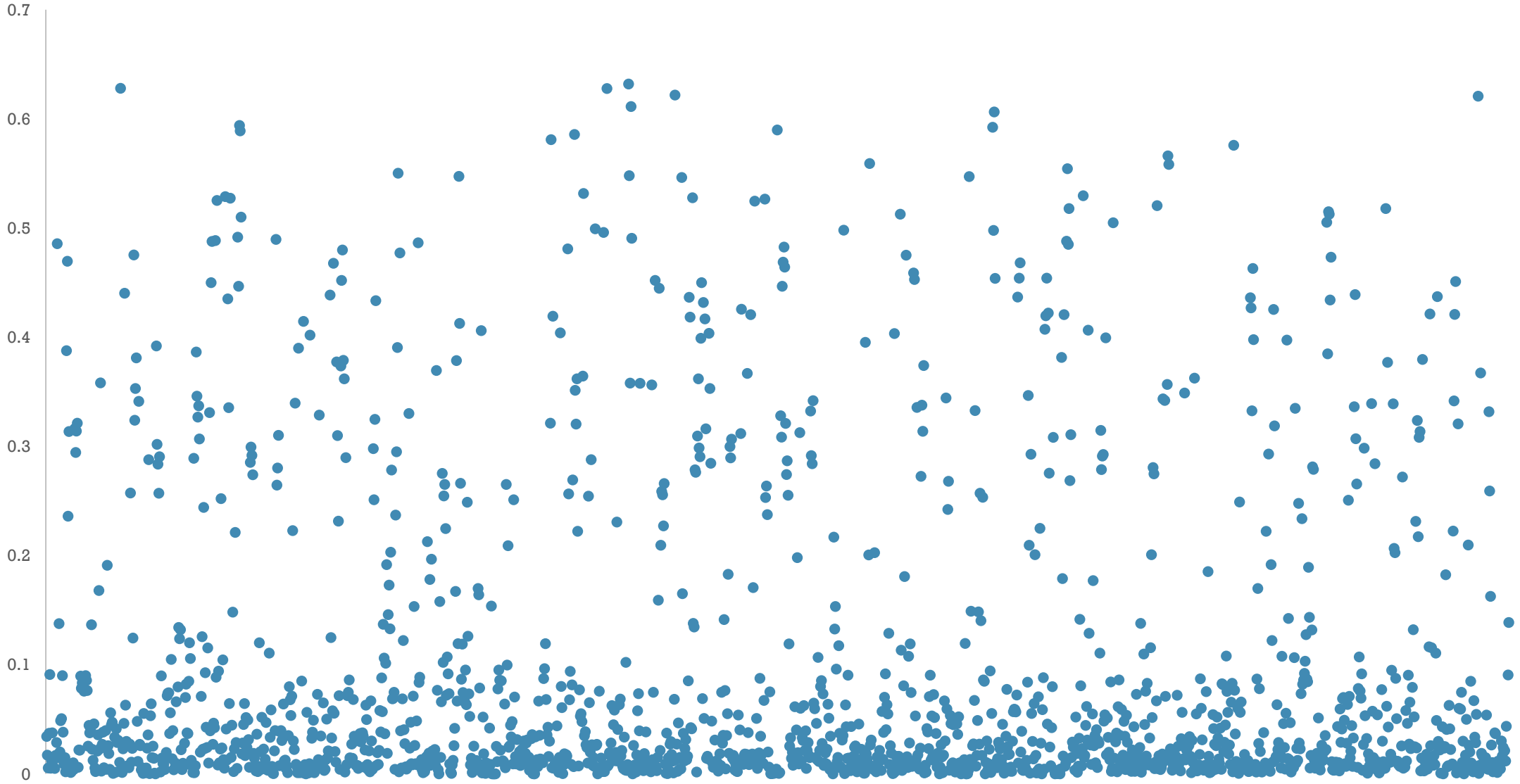
Based on histogram analysis—of pixels' intensity values—of each page

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

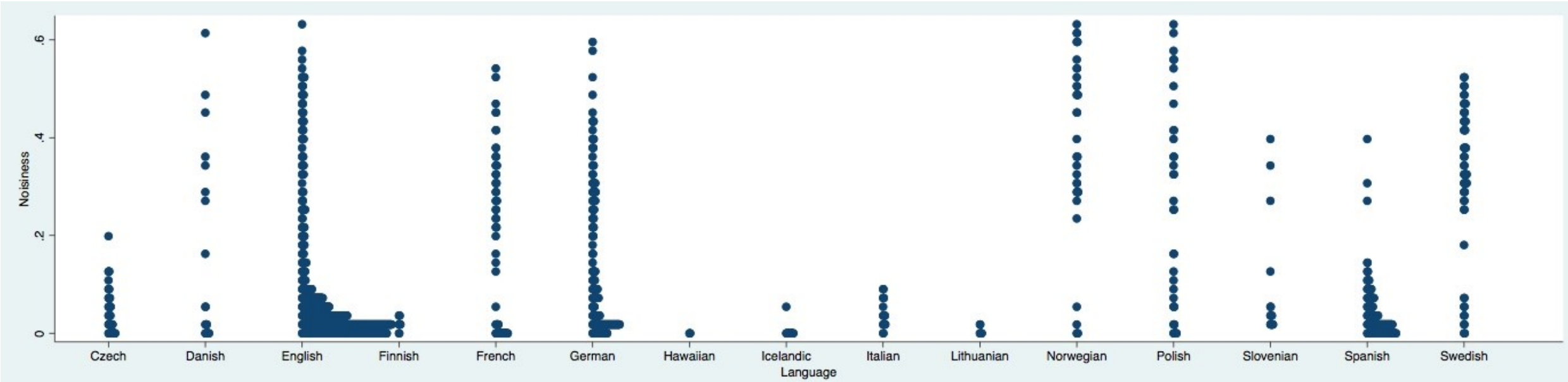
Aida

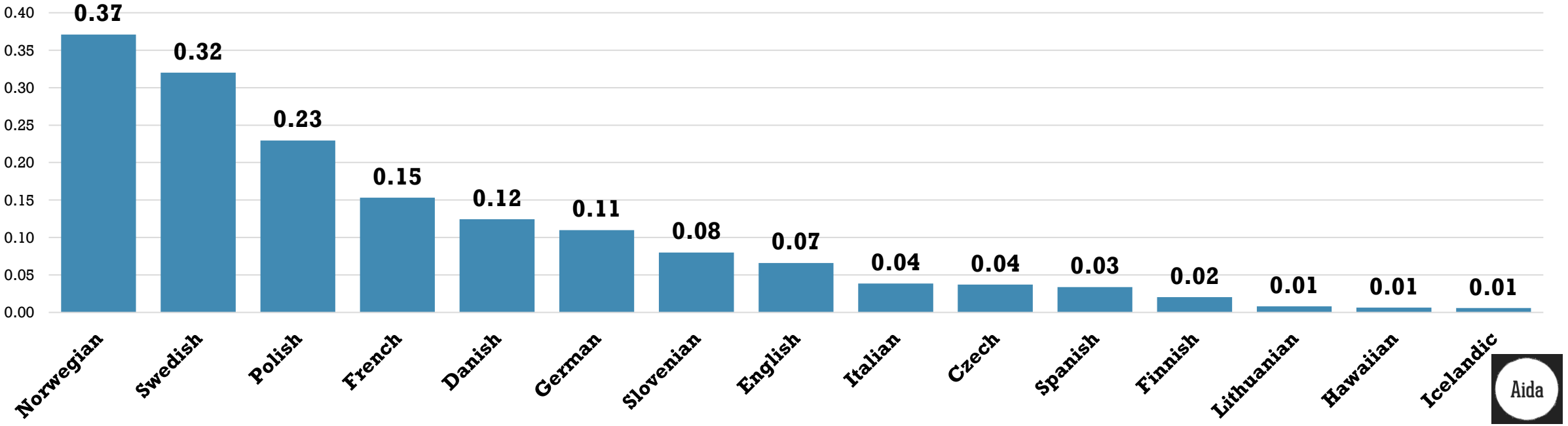
Noisiness
Test Set



Aida

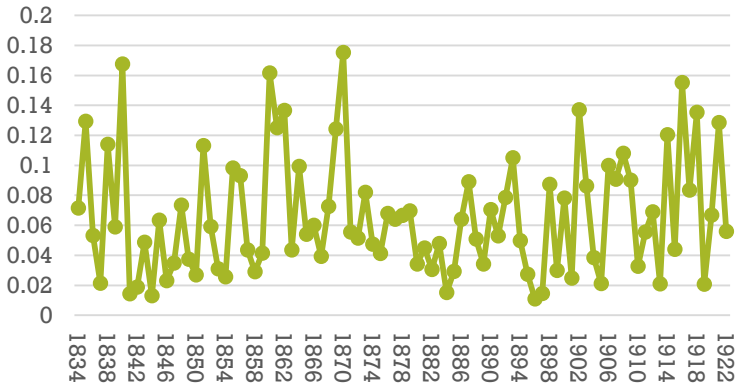
Views of Noisiness



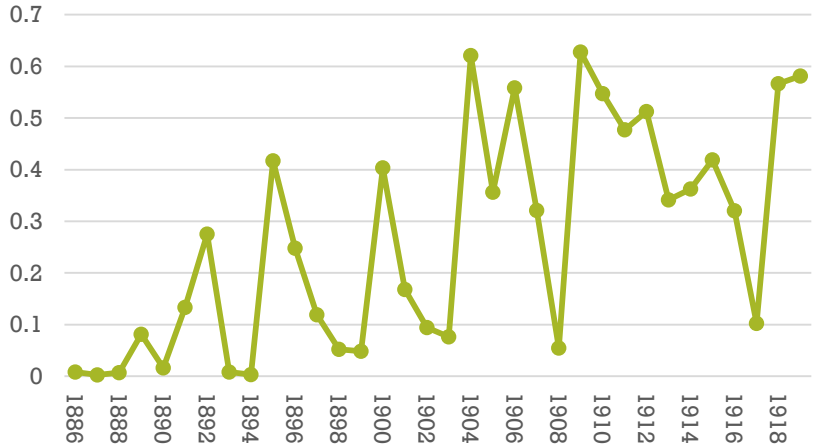


Aida

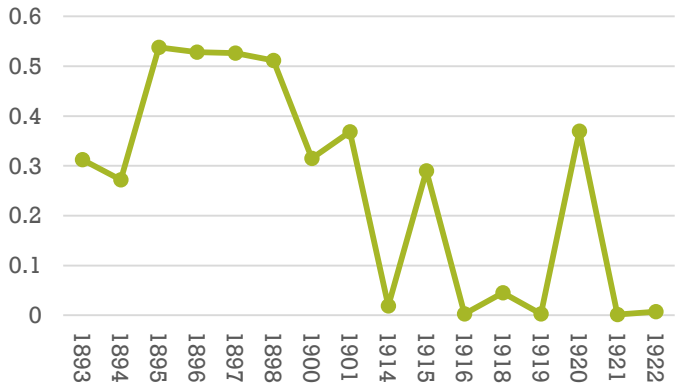
NOISE EFFECT: ENGLISH



NOISE EFFECT: POLISH



NOISE EFFECT: NORWEGIAN



Aida

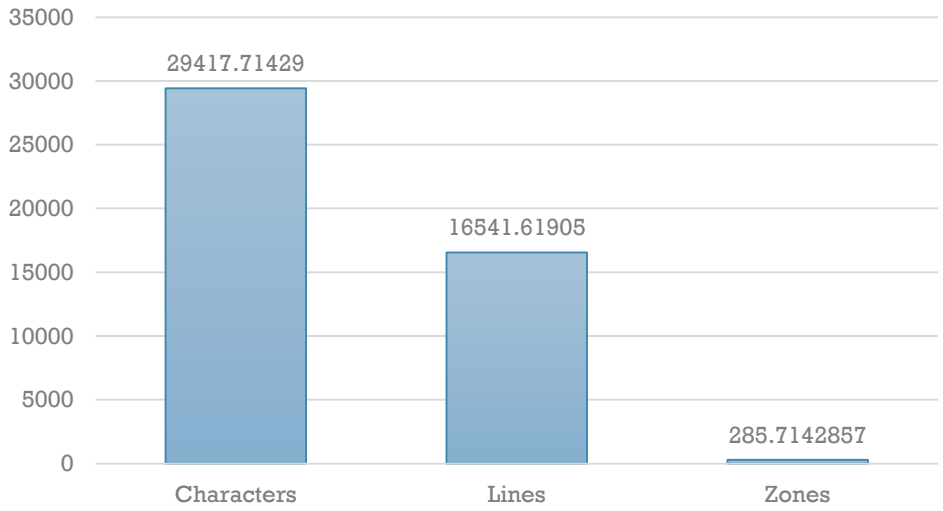
ANALYSIS: CHARACTERS, LINES, ZONES

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

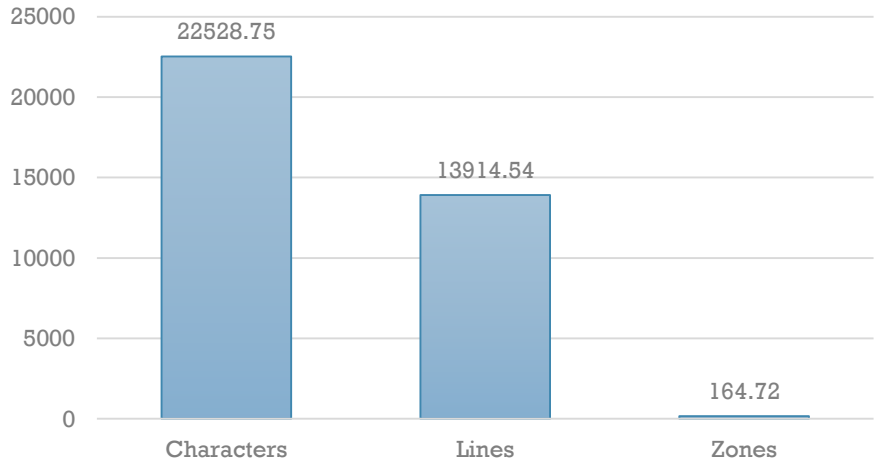
As a way to begin measuring both complexity and density of the textual content on the page

Aida

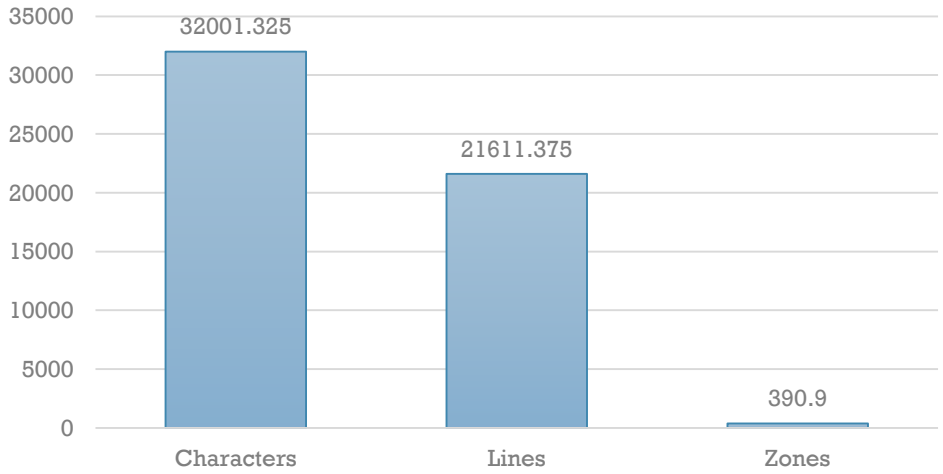
Italian



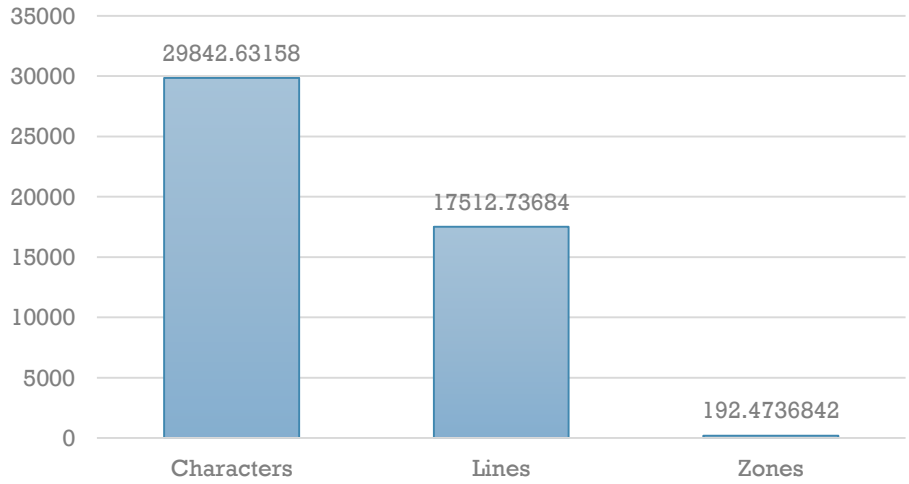
Spanish



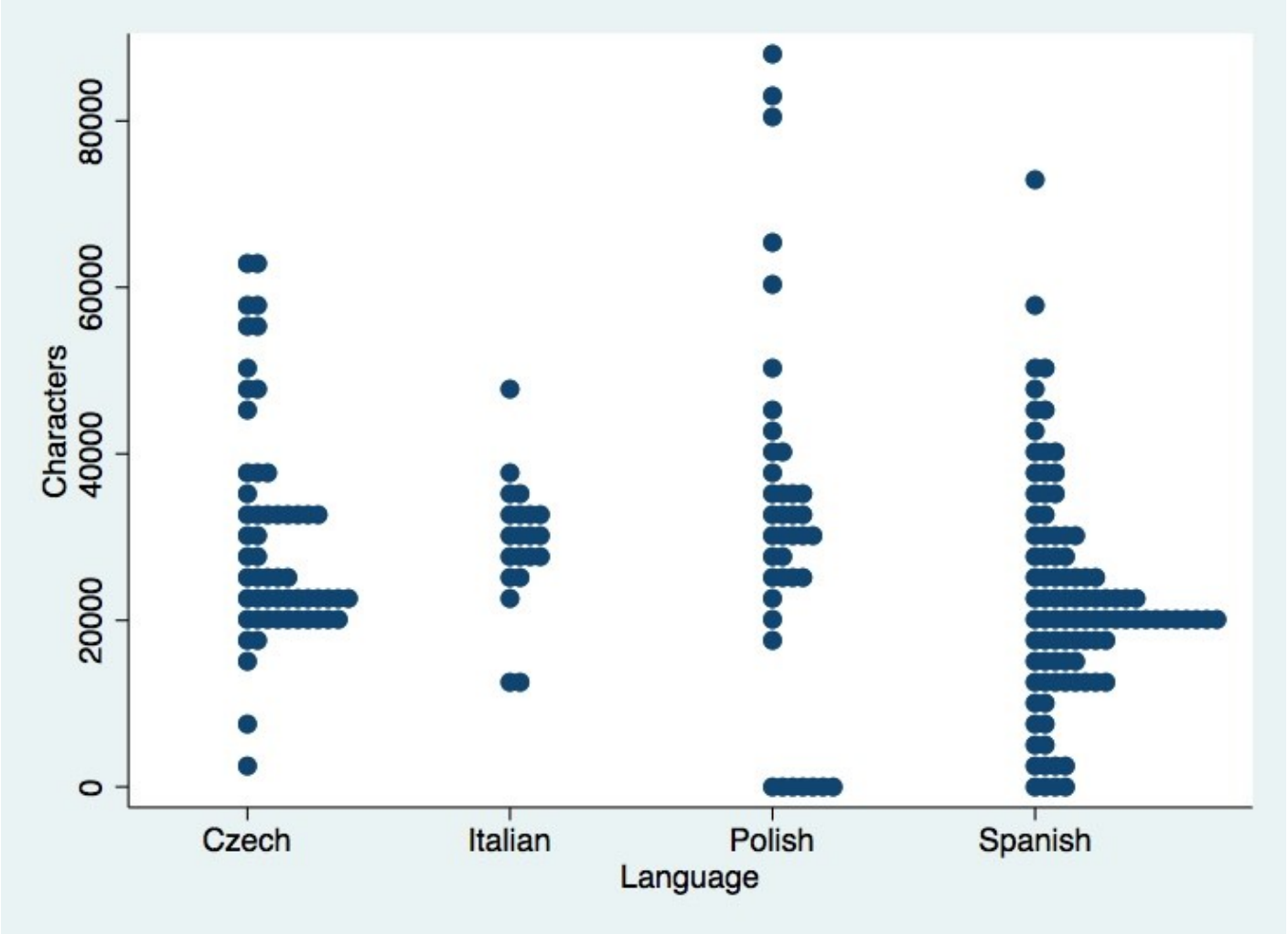
Polish

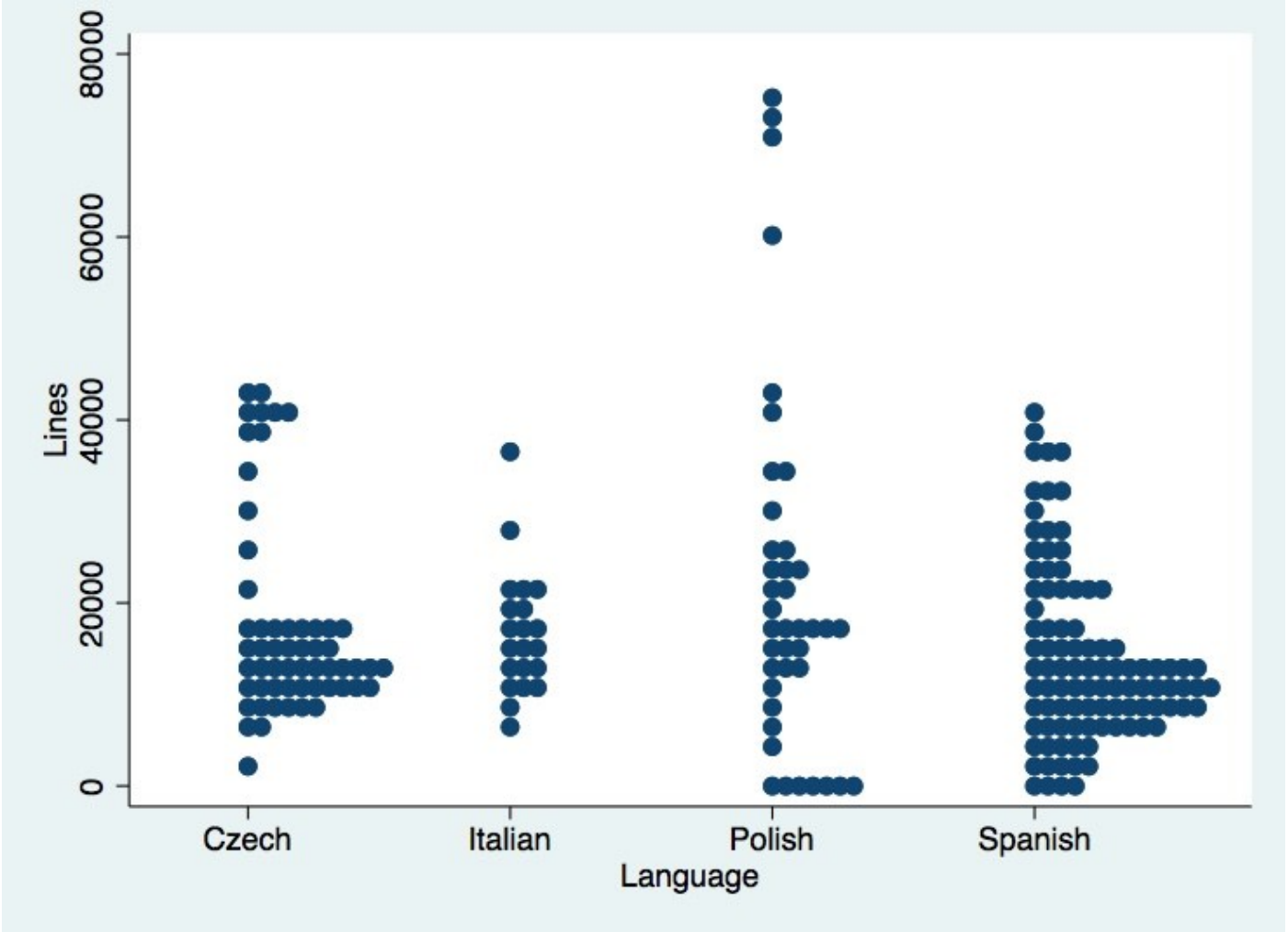


Czech

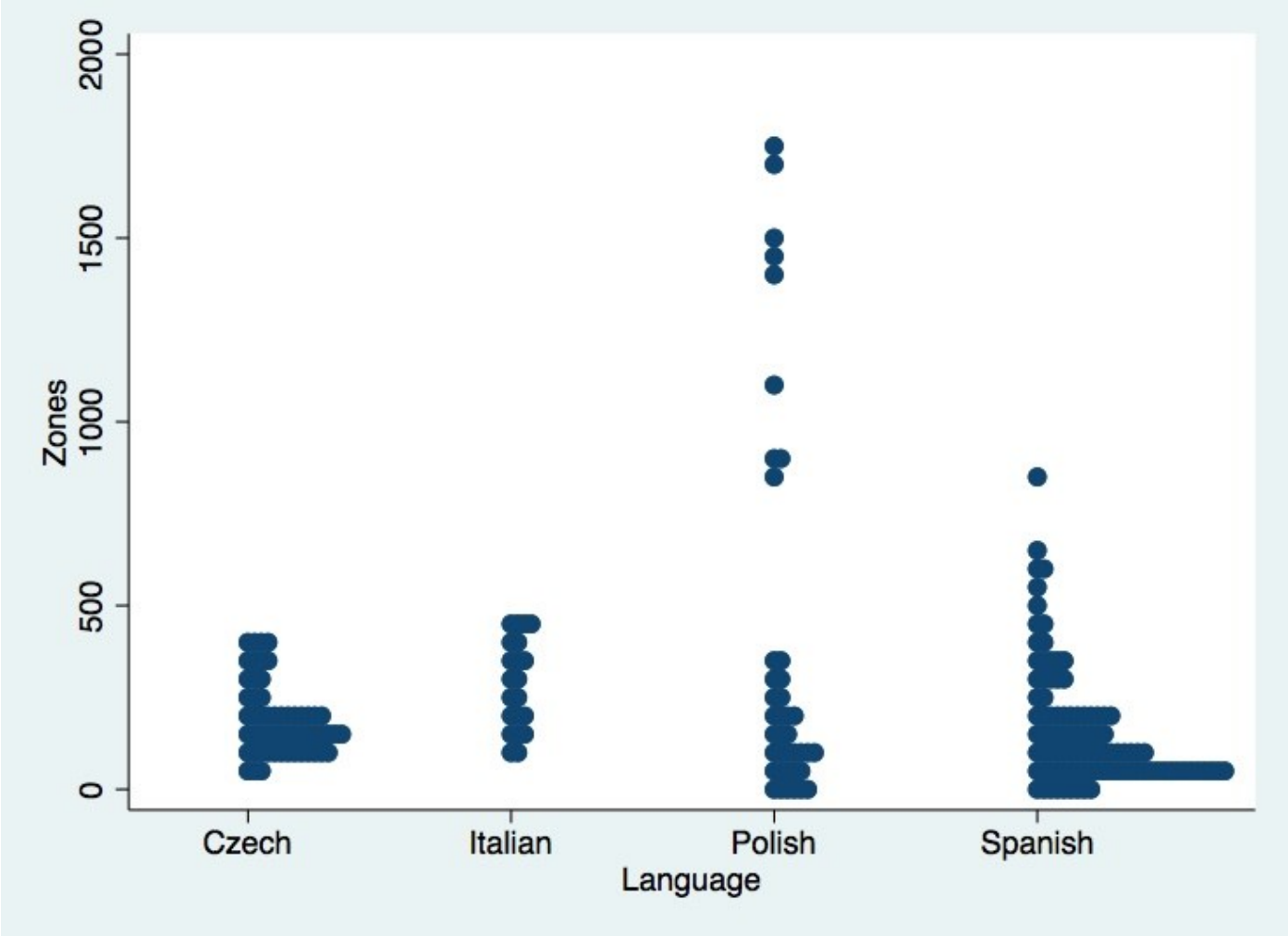


Aida


Characters
0
20000
40000
60000
80000
Czech
Italian
Polish
Spanish
Language


Lines
0
20000
40000
60000
80000
Czech
Italian
Polish
Spanish
Language

Aida


Zones
0
500
1000
1500
2000
Czech
Italian
Polish
Spanish
Language

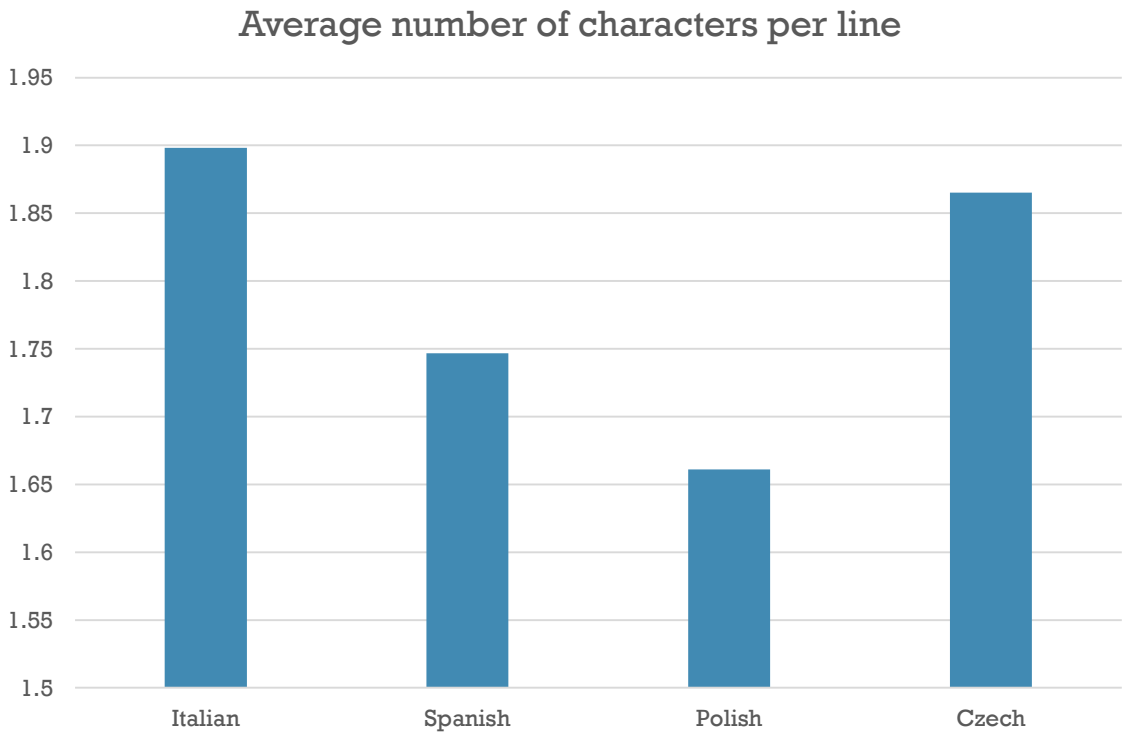
ANALYSIS 2: CHARACTERS, LINES, ZONES

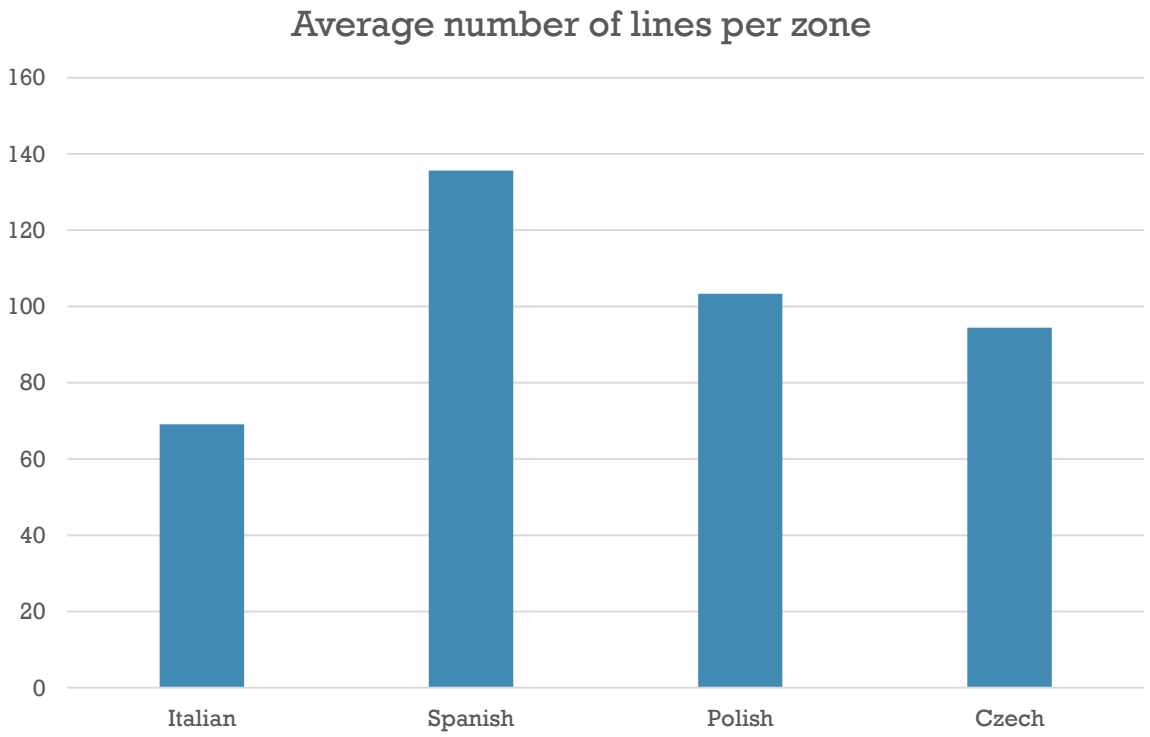
Numbers of characters, lines, and zones vary across newspapers of different languages

- 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

Aida


Average number of characters per line
1.95
1.9
1.85
1.8
1.75
1.7
1.65
1.6
1.55
1.5
Italian
Spanish
Polish
Czech


Average number of lines per zone
160
140
120
100
80
60
40
20
0
Italian
Spanish
Polish
Czech

Aida

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

Aida