Early Modern OCR Project (eMOP) at Texas A&M University: Using Aletheia to Train Tesseract

Katayoun Torabi, PhD Student in English
Bryan Tarpley, PhD Student in English
Dr. Jessica Durgen, Post-Doctoral Fellow (not present)
The Problem:

- The IDHMC is tasked with performing OCR on over 45,000,000 page images of historical documents which were scanned using poor resolution scanners with problems such as idiosyncratic fonts and bleed-through.

- In-house, we have at our disposal 2 programmers and ~ 5 part-time students in the English department and the Digital Humanities. We are also privileged to work with various collaborators, including PRImA.

- We may only work with open-source solutions (no ABBY FineReader).

- The time frame for our project is approximately 2 years.
Historical Documents

And great honour to thee is given,
In the heavenly habitation:
Where the fruit of thy womb is living,
For thy alms become a monument.

Of thy fruit art thou blessed, O glorious
Lady,
For the fruit of life we received by thee,
O blessed mother of God, Virgin Mary.

God, which with double tide hath fulfilled the most blessed virgin grace, as
well in the conception, as in the birth of the right entirely beloved sonne (her vir-
ginite saucd) which also hath multiplied her Loves, though the resurrection of her
famed sonne. Cannot we believe thee, that
by her merites and prayer, we may attain
to the unparallel love, whereby the being
as in the death now cloth with thee in hea-
ven for evermore. By Christ our Lord. A
men. 

Close O source of virgines all,
In thine honor and grace special,
Receiving a thousands tidings
The principalities of angels
impossible.

And in the dignitate of saints esteemed,
More then can be said. O, O, clothe
Ligatures
Germanie
And because it hath much heat, it expelleth what soever is unclean, and therefore restoreth the natural heat. For age is nothing else than a lessening and diminishing of natural heat, which is therefore diminished, because the moving is hindered as I laide offire; for there is like reason of this unto that. Moving is hindered, because of abundance of earthy matter, because the earth hath very muche matter, and is destitute of all moving. Therefore that water being of so temperate a heat, it shall neither be the hearte, nor dissolveth the liver; wherefore this onlye can perform the things that we have spoken. When burning water therefore or Aquavitæ retineinge his purity and subtil matter: by the longe motion of circulacion, hath put of and rid away his heat and Sharpnes of taste and smell, it is became of the
Other Special Characters

• Other special characters encountered are letters with suspension marks, rotunda r’s, and letters with superscripts.
Long ‘s’ in the word ‘greatnesse’
Long ‘s’ and ‘f’

So high or low, dost raise thy formall hat:

‘s’ with a half cross-bar
‘f’ with a full cross-bar
PRImA, the developers of Aletheia Desktop
TO THE MOST REVEREND FATHER in God, and his singular good Lord, Edmond, by the grace of God, Archbishop of Canterbury, Primate and Metropolis of all England, Thomas Tusser, with the pleasant verses of the Spirit in Christ Jesus.

For long exercise in translating such Latine Commentaries upon the holy Scriptures, as I thought most like to further my countrymen, which understand not that tongue, to the solid knowledge of true Religion: at last I took in hand M. Caluins exposition upon Saint Pauls Epistles to the Corinthians. And, as in my poor judgment, the writer is a most excellent instrument of God, for the simple setting forth of his truths, so in making my choice...
Step 1: Upload the TIFF image in Aletheia
Step 2: Binarize the image
Step 3: Remove noise
Step 4: Running the automatic segmentation tool in Aletheia

A long exercise in translating such Latine Commentaries upon the holy Scriptures, as I thought most like to further my country men, which under
TO THE MOST REVEREND FATHER IN GOD, AND HIS SINGULAR GOOD LORD, EDMOND, BY THE GRACE OF GOD, ARCHBISHOP OF CANTERBURY, PRIMATE AND METROPOLITAN OF ALL ENGLAND, THOMAS TIMON, WHOSE THE PLENTIFUL VIGORS OF THE SPIRIT IN CHRIST JESUS.

FTER LONG EXERCISE IN TRANSLATING SUCH LATINE COMMENTARIES UPON THE HOLY SCRIPTURES AS I THOUGHT MOST LIKE TO FURTHER MY COUNTRY.
TO THE MOST REVEREND FATHER IN GOD, AND HIS SERVANT, THE GOOD LORD EDMUND, BY THE GRACE OF GOD, ARCHBISHOP OF CANTERBURY, PRIMATE AND METROPOLITAN, FROM THE ENGLAND, THOMAS SIMON, WRITING IN THE PLURALITY OF THE SPIRIT IN CHRIST.

FERILD EXERCISE IN TRANSLATING SUCH LATIN COMMENTARIES UPON THE BODY OF SCRIPTURES, AS I THOUGHT MOST LIKE TO FURTHER MY COUNTRY.
To the most reverend Father in God, and my singular good Lord Edmond, by the grace of God, Archbishop of Canterbury, Cambridge and Metropolis of all England, Thomas Linne, a minister of the plentiful riches of the Spirit, in Christ Jesus.

Even long experience in translating such Latin Commentaries upon the holy Scriptures, as I thought most likely to further my country men, which under
The segmentation tool generates text for each character and a corresponding XML file
To the Most Reverend Father in God and to the Right Reverend Father in Christ.

Guler good Lord, I have by the tene of all England, wished the plentiful riches of the Spirit, in Christ Jesus.

Further long exercise in translating such Latine Commentaries upon the holy Scriptures, as if thought most like to further my country.
Misreads

- Ligature boundaries are sometimes confused
Step 5: Glyph Correction

• The parameters of a glyph may be corrected using the edit button to adjust lines
Step 6: Text Correction

• There is difficulty reading letters that are distorted, obscured by noise or ink bleed-through, or if they are especially faint.

![Image of text correction interface]
Step 6: Text Correction

- Text misreads must be corrected by typing in the correct text in the Text Content Box for each corresponding character:
Font Sets

To date the Aletheia team at the IDHMC has produced fourteen font training sets

<table>
<thead>
<tr>
<th>Font Name</th>
<th>Publisher</th>
<th>Publication Year</th>
<th>Images Processed</th>
</tr>
</thead>
<tbody>
<tr>
<td>JDAY1559_001</td>
<td>John Daye</td>
<td>1559</td>
<td>15</td>
</tr>
<tr>
<td>MFLE1633_94_017</td>
<td>Miles Flesher</td>
<td>1633</td>
<td>11</td>
</tr>
<tr>
<td>TCOT1632_82_001</td>
<td>Thomas Cotes</td>
<td>1632</td>
<td>11</td>
</tr>
<tr>
<td>IJAGG_guyot_83_99</td>
<td>François Guyot</td>
<td>Unknown</td>
<td>13</td>
</tr>
<tr>
<td>BRA_2788_002</td>
<td>Pierre Haultin</td>
<td>Unknown</td>
<td>8</td>
</tr>
<tr>
<td>cbil1692_110_002</td>
<td>Charles Bill</td>
<td>1692</td>
<td>6</td>
</tr>
<tr>
<td>cbil1693_116_001</td>
<td>Charles Bill</td>
<td>1693</td>
<td>10</td>
</tr>
<tr>
<td>cbil1702_116_1_001</td>
<td>Charles Bill</td>
<td>1702</td>
<td>14</td>
</tr>
<tr>
<td>hhi1686_82_1_001</td>
<td>Henry Hills</td>
<td>1686</td>
<td>15</td>
</tr>
<tr>
<td>jbill1628_86_001</td>
<td>John Bill</td>
<td>1628</td>
<td>15</td>
</tr>
<tr>
<td>IJAGG_guyot_83_15_B_ italics</td>
<td>François Guyot</td>
<td>Unknown</td>
<td>12</td>
</tr>
<tr>
<td>BRA_2788_004_B_ italics</td>
<td>Pierre Haultin</td>
<td>Unknown</td>
<td>5</td>
</tr>
<tr>
<td>MFLE1633_94_017_B_ italics</td>
<td>Miles Flesher</td>
<td>1633</td>
<td>11</td>
</tr>
<tr>
<td>TCOT1632_82_001_B_ italics</td>
<td>Thomas Cotes</td>
<td>1632</td>
<td>14</td>
</tr>
</tbody>
</table>
Training Tesseract using Aletheia

• Once the font training libraries were in place, we started to train Tesseract by running each prepared font set through the Tesseract engine so that it would learn to read and recognize the characters of those fonts.
Aletheia and Tesseract

Aletheia Glyph Identification

Tesseract Box File

Accurate OCR Output
Roadblocks:

• How do we “feed” Tesseract only pristine instances of each letter for training purposes?

• How do we ensure that each glyph has been identified correctly and consistently?

• How do we eliminate the misidentification of pixel artifacts as characters?
him, would often rub his Hands, and cry, Sprage, Sprage, they follow us still. I am very sensible later Times have not been over favourable in their Sentiments of that unfortunate Prince’s Valour, yet I cannot omit the doing a Piece of justice to his Memory, in relating a Matter of Fact, of which my own Eyes were Witnesses, and saying, That if Intrepidity, and undauntedness, may be reckoned any Parts of Courage, no Man in the Fleet better deserv’d the Title of Courious, or behav’d himself with more
Promising Results

Before Franken+

OB I UxjnR yOU; OWOUM hcz NUUC wcxc GUWY
EUC OUCC kCFoz-C 1 FOICC CO ckP pUX-pzzaFPP wkc* 0
NU*ICC Uzc hzuFfc 2 1 10U8~

278
VVRC Lzfly TC hcc L0xch YOUPIC *y 2
POB- N#M#UXCUC-
ffNp N2y' kUC yOU w11?

Inff- I U22y UOC vCKCIy-
HCp VcScly?

YOU FUC UU Offwjck IRkcx VOWCY hU$ L-
ThQUBh yOU WOUUM fCCIC C9vFphxlc chc SC2$$ wflck Or'k#
SkOUM ycc f2y' SfuO 80IU8; VCzCIy

After Franken+

That when thou meet'af6 one, with enqtziring eyes
Doit fearch, and like a needy broker ptize

The Glke, and gold he weares, and to that rate
Sohigh or low, doG: raife thy formall hat:

That wilt confort none, Untill thou have knowne
What lands hee hath in hope, or of his owne,

As though all thy companions Ishould make thee
gointntes,and marry thy deare company.

Why HaouEd'R thou that dofk Uot onely apptove;
But in ranke itchie luf?, deGre, andlove

The nakednefe and barrenneHe to enjoiy,
Thank You!

- Kathy Torabi: torabik@tamu.edu
- Bryan Tarpley: bptarpley@tamu.edu
- The IDHMC: http://idhmc.tamu.edu
- The EMOP Project: http://emop.tamu.edu