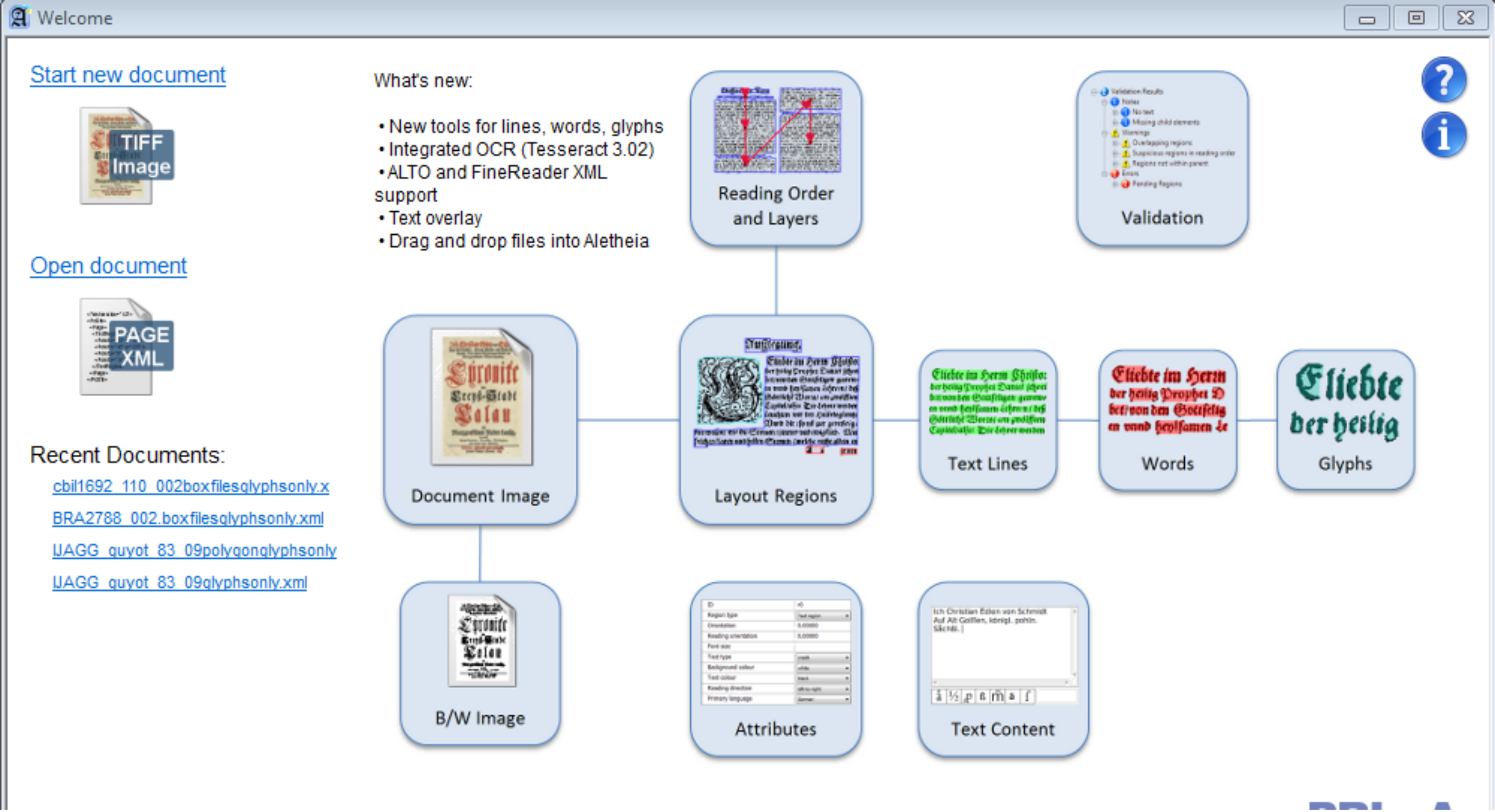
EMOP

Early Modern OCR Project



What's new:

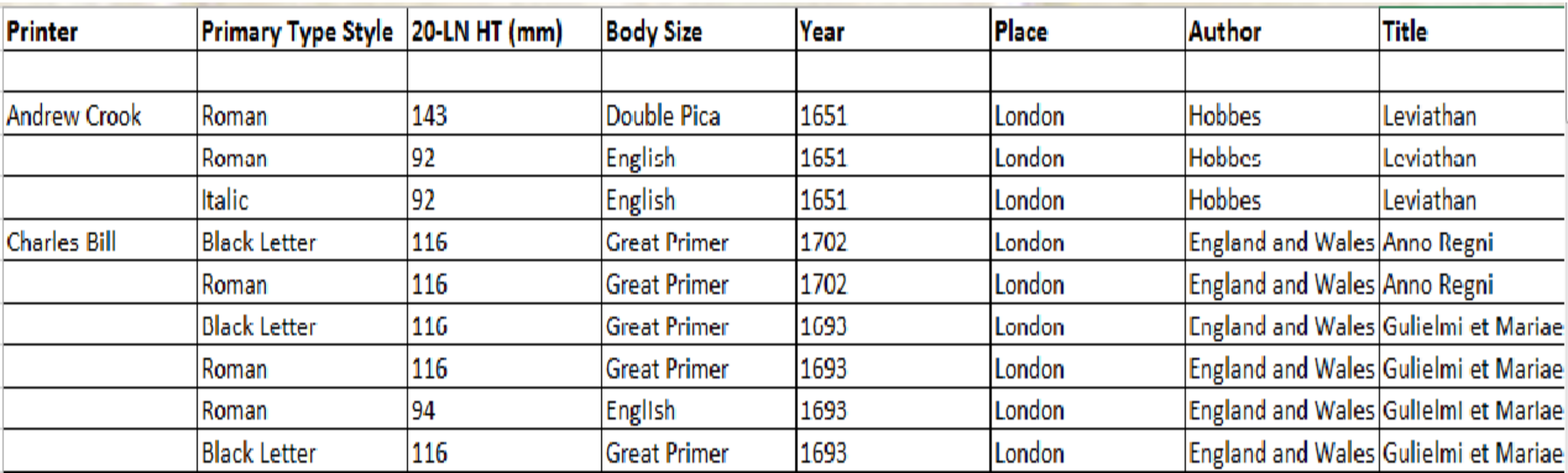
- New tools for lines, words, glyphs
- Integrated OCR (Tesseract 3.02)
- ALTO and FineReader XML support
- Text overlay
- Drag and drop files into Aletheia

Recent Documents:

cbil1692_110_002boxfilesglyphsonly.x
BRA2788_002.boxfilesglyphsonly.xml
IJAGG_guyot_83_09polygonglyphsonly
IJAGG_guyot_83_09glyphsonly.xml

Step One: Naming Font Sets

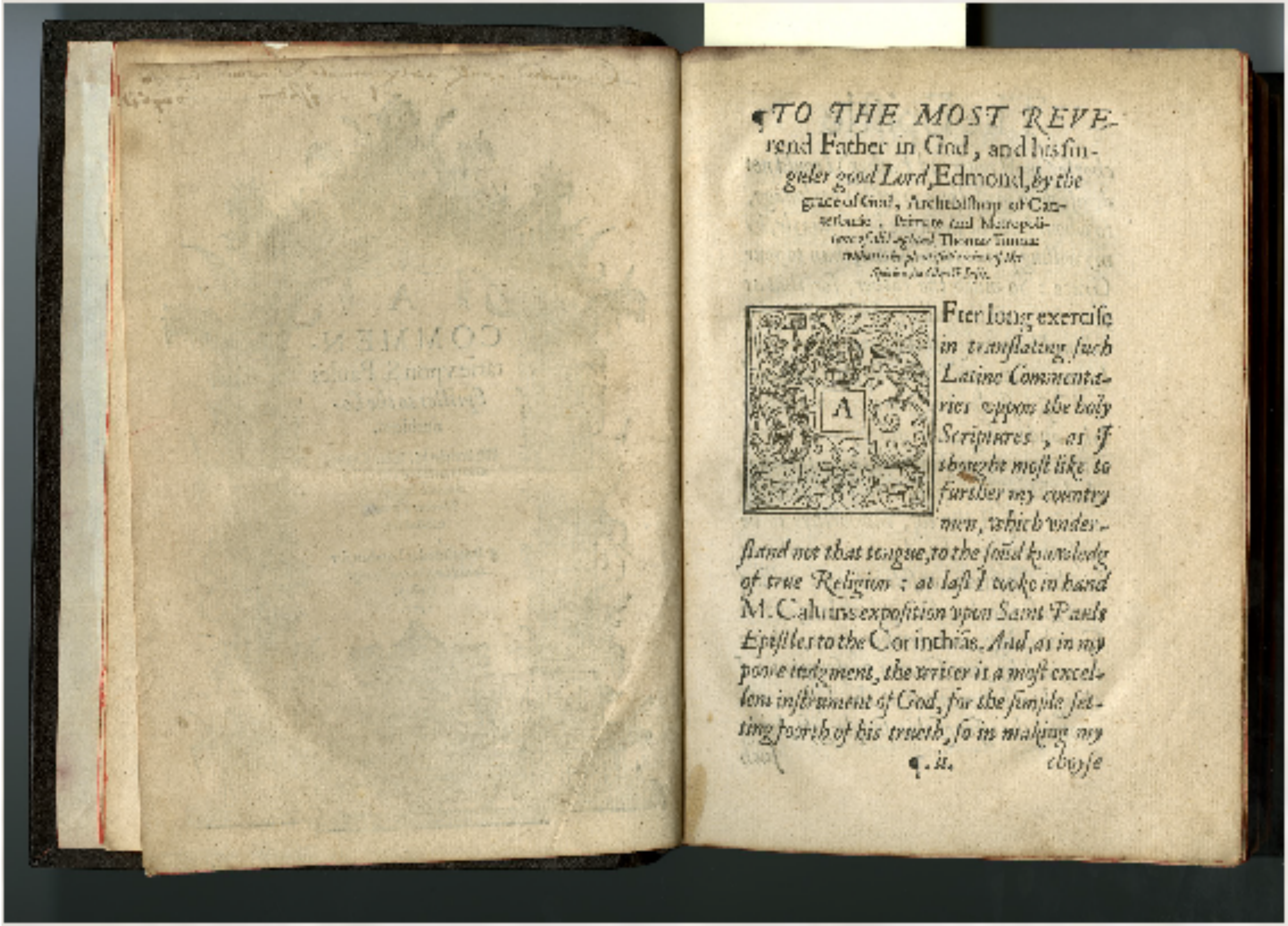
Page images are named according to publisher, publication year, and the 20-line height. For example, a text image from the 1702 *Anno Regni text would be named cbil1702_116.*



Printer	Primary Type Style	20-LN HT (mm)	Body Size	Year	Place	Author	Title
Andrew Crook	Roman	143	Double Pica	1651	London	Hobbes	Leviathan
	Roman	92	English	1651	London	Hobbes	Leviathan
	Italic	92	English	1651	London	Hobbes	Leviathan
Charles Bill	Black Letter	116	Great Primer	1702	London	England and Wales	Anno Regni
	Roman	116	Great Primer	1702	London	England and Wales	Anno Regni
	Black Letter	116	Great Primer	1693	London	England and Wales	Gulielmi et Mariae
	Roman	116	Great Primer	1693	London	England and Wales	Gulielmi et Mariae
	Roman	94	English	1693	London	England and Wales	Gulielmi et Mariae
	Black Letter	116	Great Primer	1693	London	England and Wales	Gulielmi et Mariae

Step Two: Page Segmentation and Analysis

Layout regions, lines, words, and individual glyphs in a text are identified and defined using Aletheia Desktop, a tool developed by PRImA Lab at the University of Salford that performs page segmentation and text recognition, among other tasks.



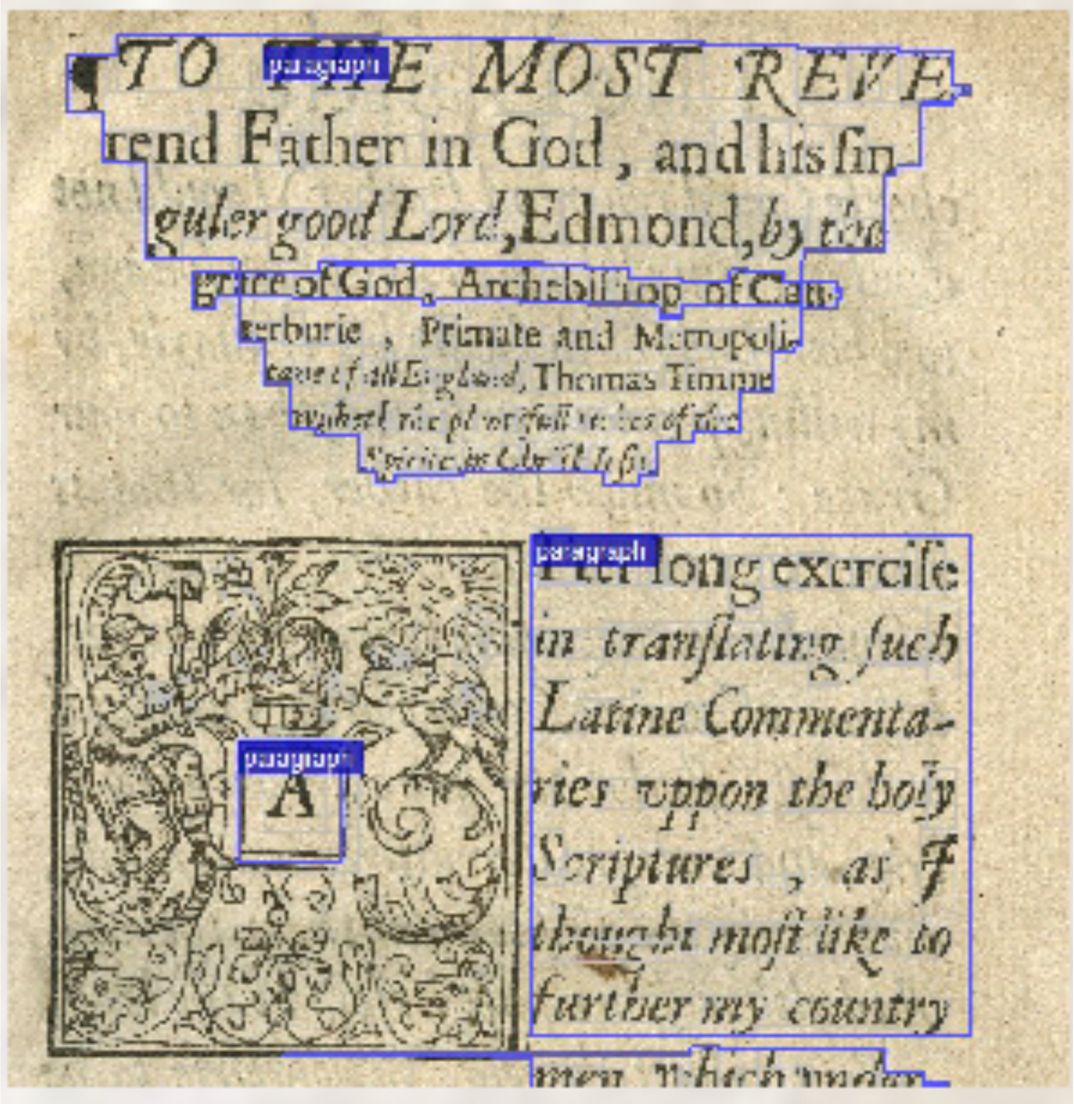
¶TO THE MOST REVE-
rend Father in God, and his ſin-
guler good Lord, Edmond, by the
grace of God, Archbiſhop of Can-
terburie, Primate and Metropoli-
tane of all England, Thomas Timme
miniſter wiſheth the increaſe of the
Spirite of Chriſt Jeſu.

AFter long exerciſe in tranſlating ſuch Latine Commentaries vppon the holy Scriptures, as I thought moſt like to further my countrymen, which underſtand not that tongue, to the ſound knowledg of true Religion: at laſt I tooke in hand M. Caluins expoſition vpon Saint Pauls Epiſtles to the Corinthiās. And, as in my poore iudgment, the writer is a moſt excellent inſtrument of God, for the ſimple ſetting forth of his trueth, ſo in making my

¶.ii. choyſe

Document Image

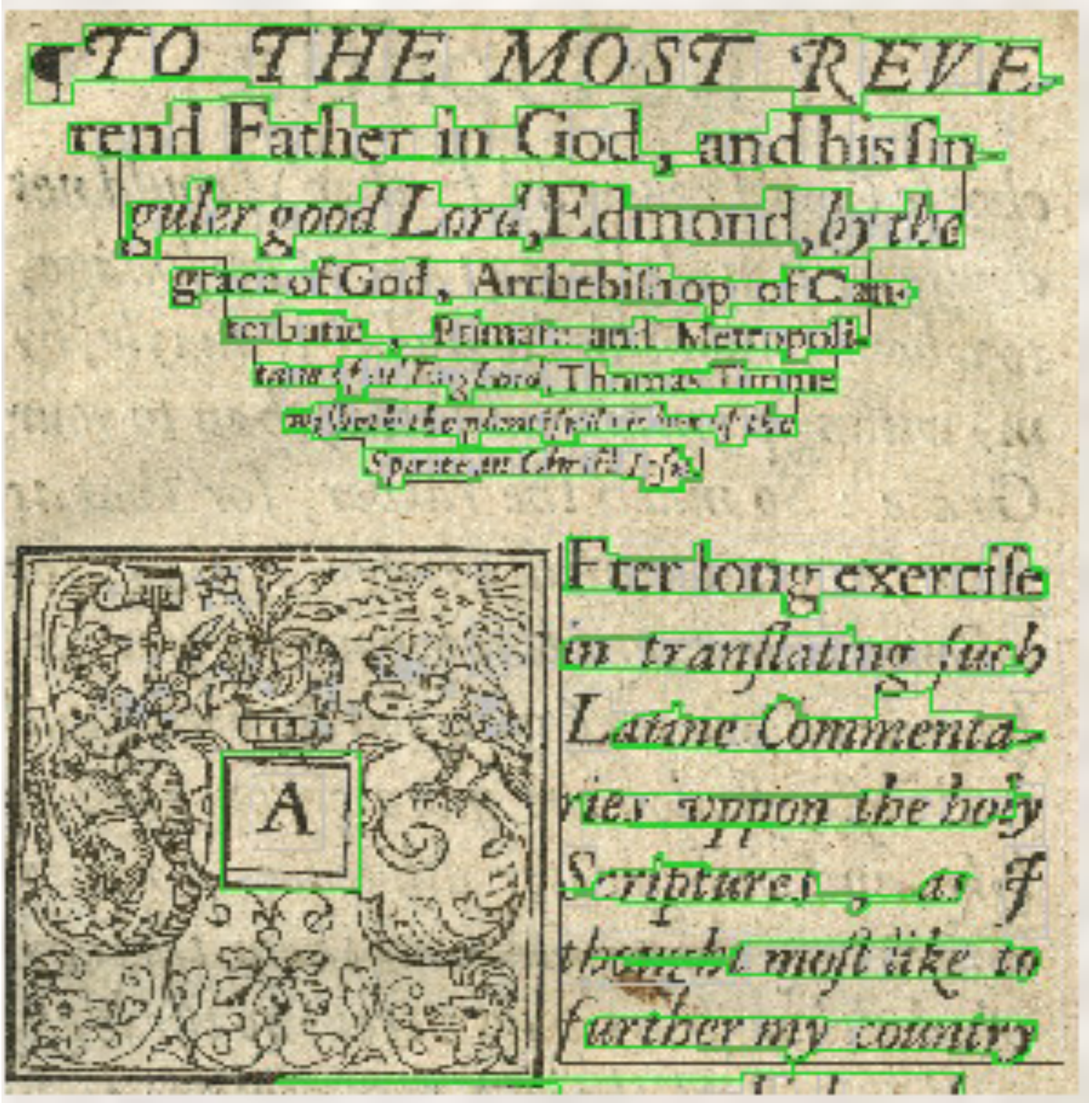
Layout Regions



¶TO THE MOST REVErend Father in God, and his ſinguler good Lord, Edmond, by the grace of God, Archebiſhop of Canterburie, Primate and Metropolitane of all England, Thomas Timme wiſheth the plentifull riches of the Spirite in Chriſt Ieſu.

After long exerciſe in tranſlating ſuch Latine Commentaries uppon the holy Scriptures, as I thought moſt like to further my country men, which under-

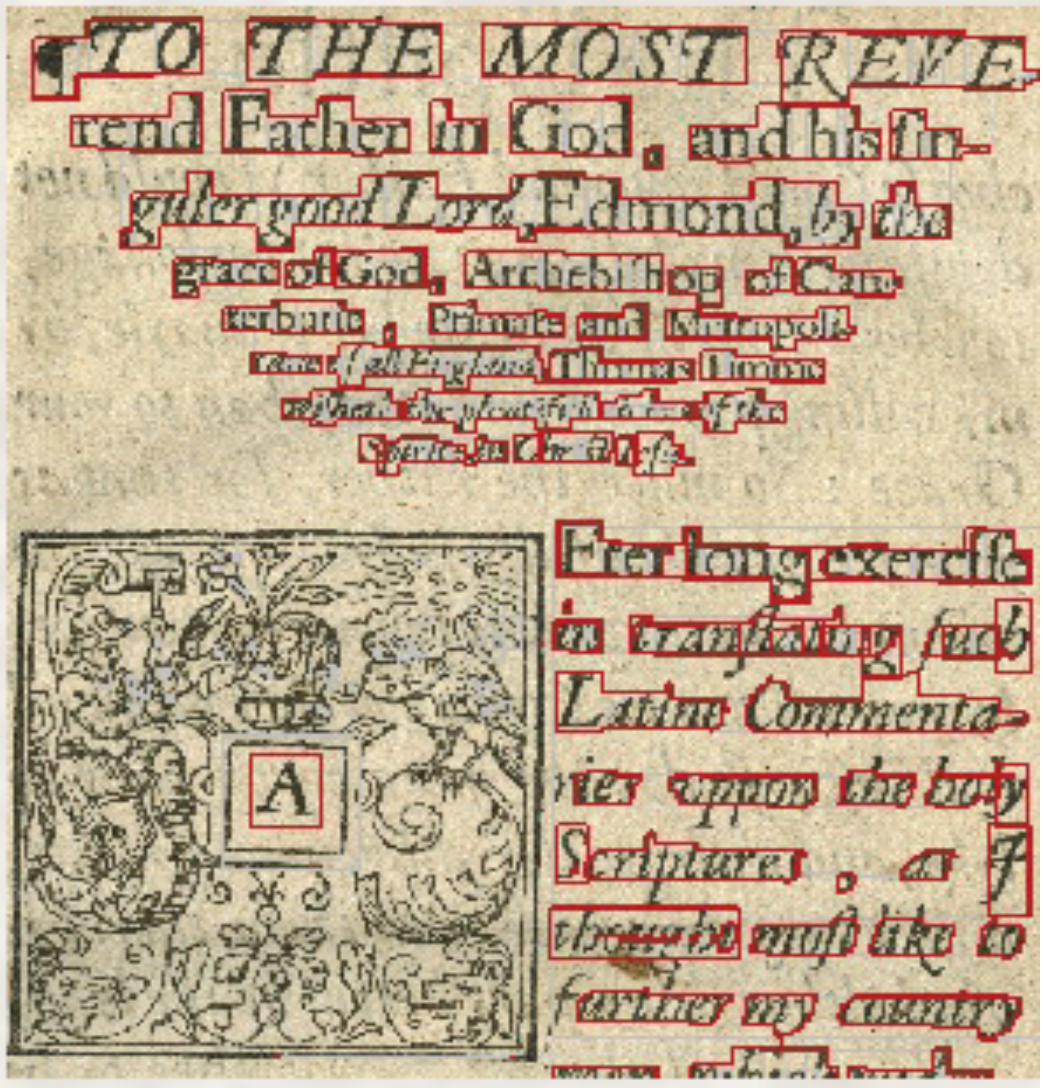
Text Lines



¶TO THE MOST REVErend Father in God, and his ſinguler good Lord, Edmond, by the grace of God, Archebiſhop of Canterburie, Primate and Metropolitane of all England, Thomas Tymme wiſheth the plentifull increaſe of the Spirite, in Chriſt Ieſu.

After long exerciſe in tranſlating ſuch Latine Commentaries vppon the holy Scriptures, as I thought moſt like to further my country

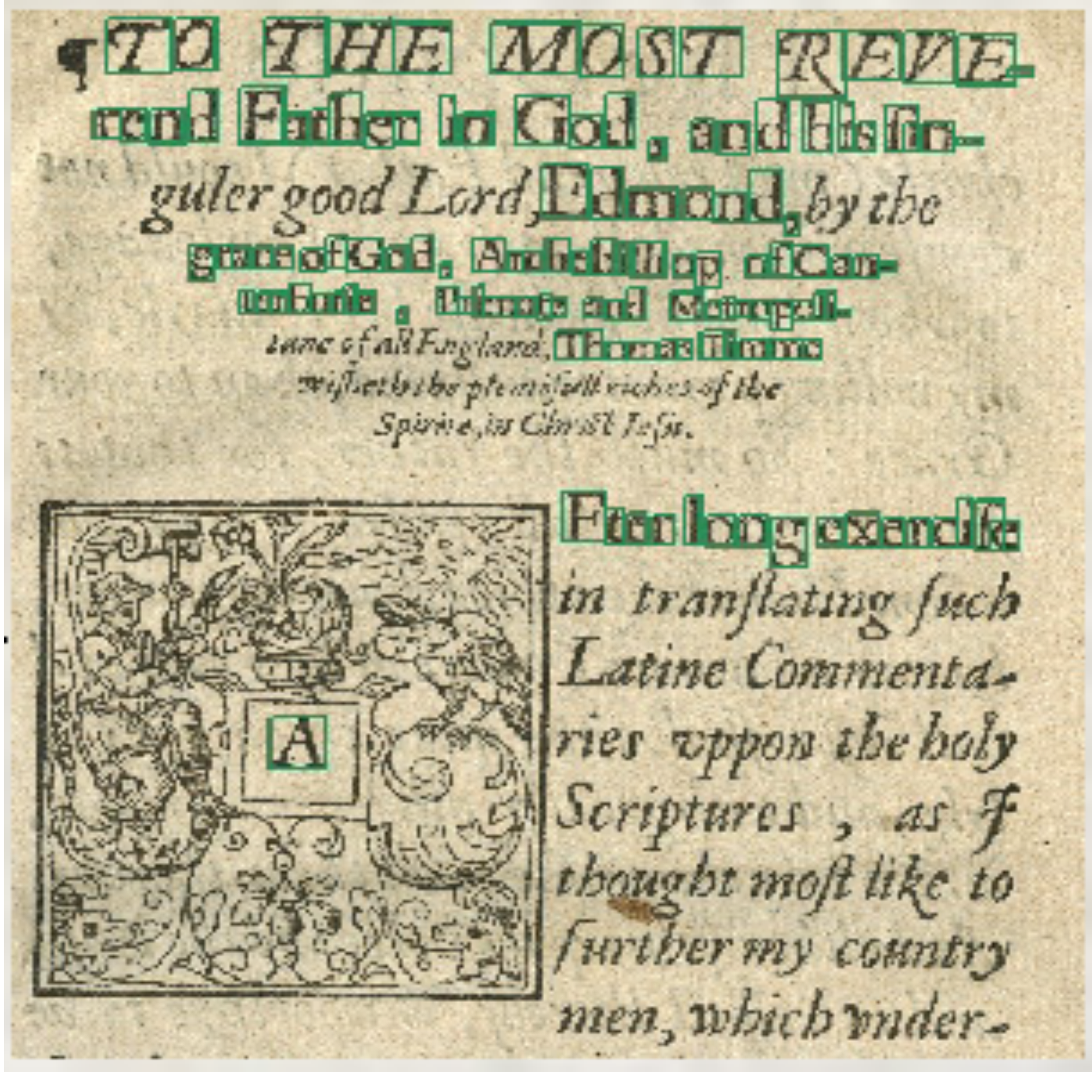
Words



TO THE MOST REVErend Father in God, and his ſinguler good Lord, Edmond, by the grace of God, Archebiſhop of Canterburie, Primate and Metropolitane of all England, Thomas Timme wiſheth the plentifull riches of the Spirite in Chriſt Ieſu.

A Fter long exerciſe in tranſlating ſuch Latine Commentaries vpon the holy Scriptures, as I thought moſt like to further my countrey

Glyphs

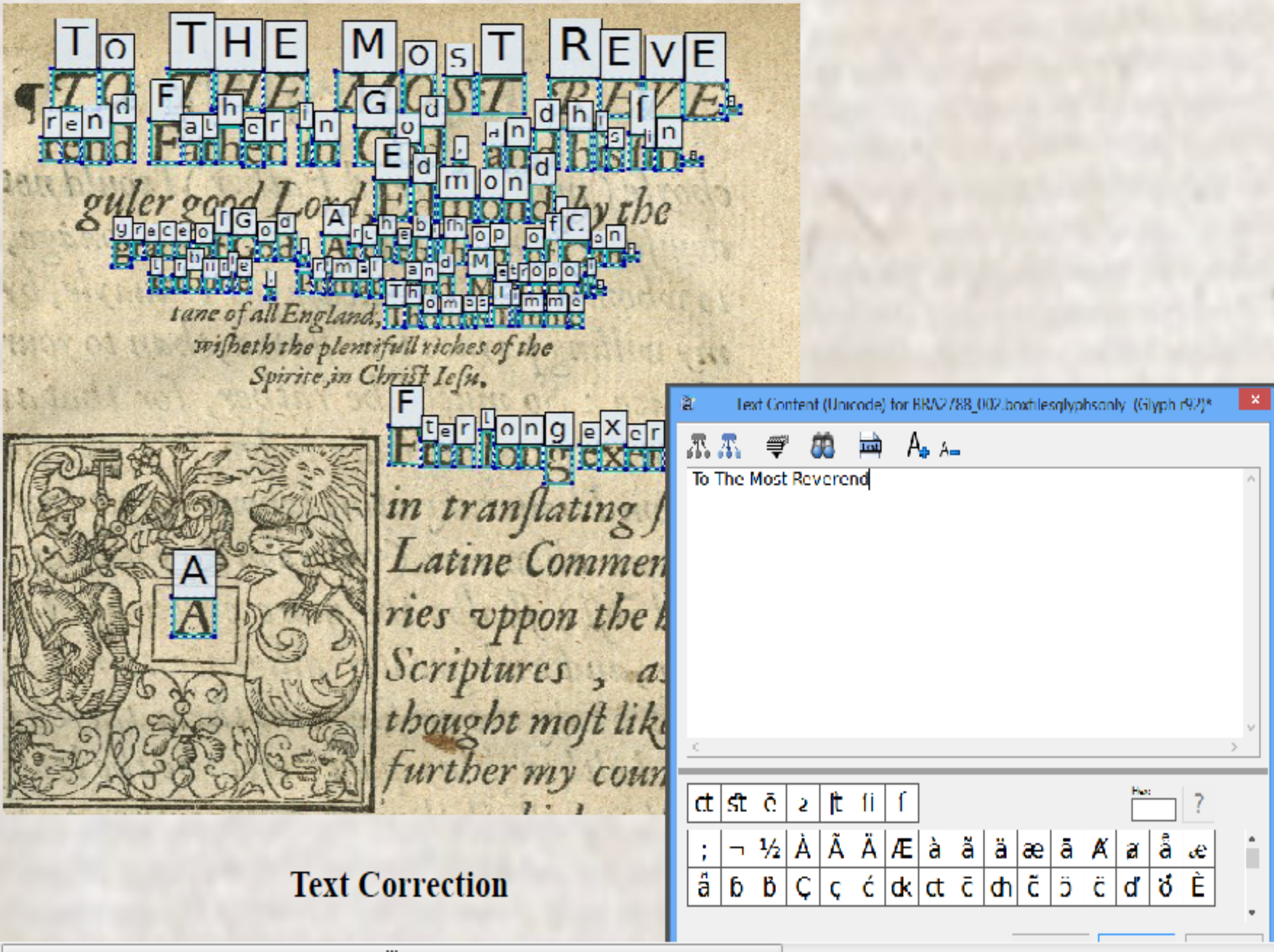


¶ TO THE MOST REVEREND Father in God, and his ſinguler good Lord, Edmond, by the grace of God, Archebiſhop of Canterburie, Primate and Metropolitane of all England, Thomas Timme wiſheth the plentifull riches of the Spirite, in Chriſt Jeſu.

After long exerciſe in tranſlating ſuch Latine Commentaries vppon the holy Scriptures, as I thought moſt like to further my country men, which vnder-

Step Three: Text Correction

Aletheia analyzes the page image and generates text for each character. Students are then responsible for fixing any "misreads" by typing in the correct text in the Text Content Box for each corresponding character--paying special attention to unique characters such as long S's, ligatures, italics, rotunda R's, suspension marks, and printer's marks.



¶ TO THE MOST REVE-
rend Father in God, and his ſin-
guler good Lord, Edmond by the
grace of God Archbiſhop of Can-
terburie, Primate and Metropoli-
tane of all England, Thomas Timme
wiſheth the plentifull riches of the
Spirite, in Chriſt Ieſu.

AFter long exper
in tranſlating ſ
Latine Commen
ries vppon the
Scriptures, a
thought moſt lik
further my coun

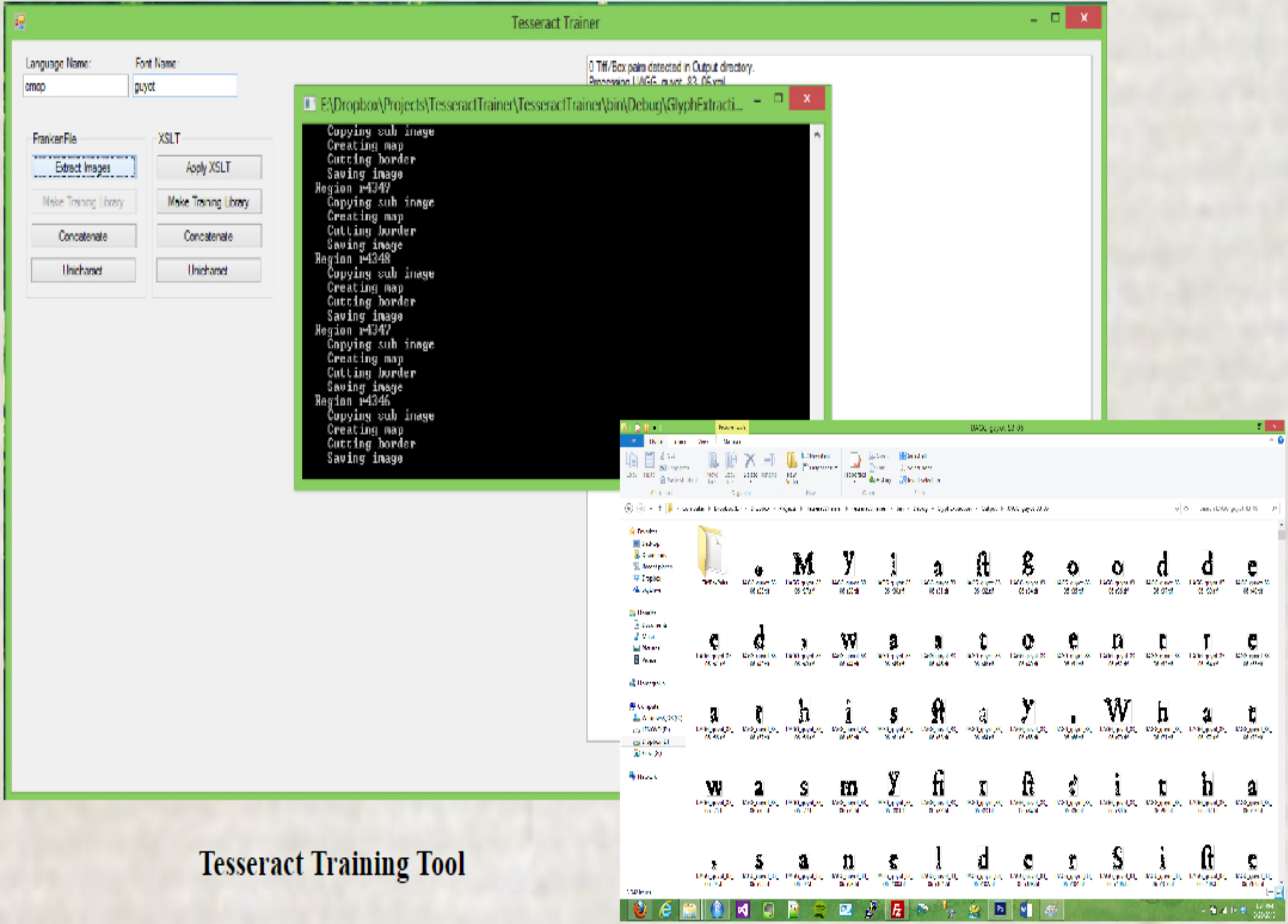
Text Content (Unicode) for BR/2788_002.boxfilesglyphsonly (Glyph r92)*

To The Most Reverend

Text Correction

Step Four: OCR Training

10-15 page images are processed for each font type and comprise a training set. A Tesseract training tool uses the font set to train the OCR engine to read and recognize that particular font.



Language Name: emop

Font Name: guyot

0 Tiff/Box pairs detected in Output directory.
Processing UAGG_guyot_83_06.xml

E:\Dropbox\Projects\TesseractTrainer\TesseractTrainer\bin\Debug\GlyphExtracti...

```
  Copying sub image
  Creating map
  Cutting border
  Saving image
Region r4347
  Copying sub image
  Creating map
  Cutting border
  Saving image
Region r4348
  Copying sub image
  Creating map
  Cutting border
  Saving image
Region r4347
  Copying sub image
  Creating map
  Cutting border
  Saving image
Region r4346
  Copying sub image
  Creating map
  Cutting border
  Saving image
```

Tesseract Training Tool