Glyph Segmentation with Aletheia - Tutorial

PRImA Research Group, University of Salford, United Kingdom

Date: 01/08/2015

Contents

1 Introduction

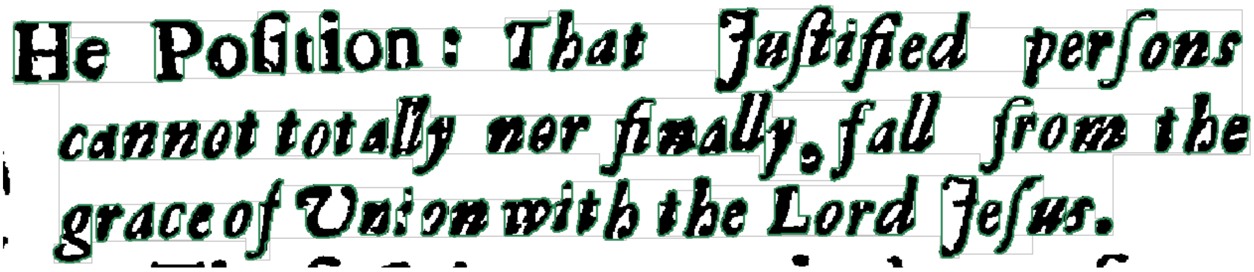
This tutorial describes different approaches to segment a document page into glyphs (character objects) using the Aletheia document analysis and ground truthing software.

Prerequisites:

- Aletheia 3.0 (see primaresearch.org/tools)
- Black-and-white document image (bitonal, preferably without noise or other artefacts)

Goal:

- Document layout description including:
  - Regions
  - Text lines
  - Words
  - Glyphs



He Poſition: That Juſtified perſons cannot totally nor finally, fall from the grace of Union with the Lord Jeſus.

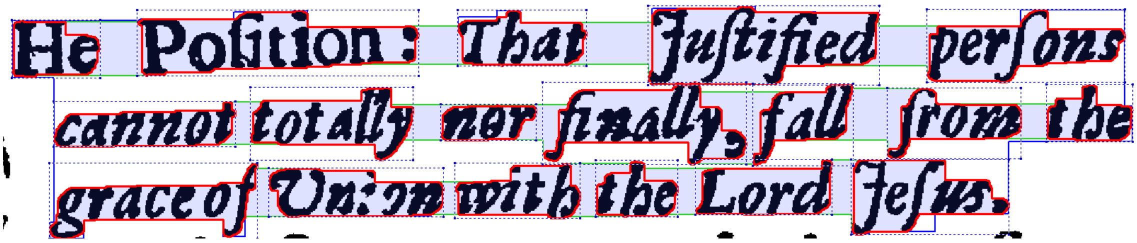
2 Workflows

There are several approaches to segment a document page into glyphs:

1. Starting from scratch
   a. Top-down: Start with regions and refine to text lines, words and finally glyphs
   b. Bottom-up: Start with glyphs and then combine to words, text lines and regions (as required)
2. Use pre-produced data: Correct segmentation results that have been produced using an external tool (e.g. Tesseract or Abbyy FineReader)

3 Top-down approach (from regions to glyphs)

- Create text regions, text lines and words according to the user guide or other tutorials

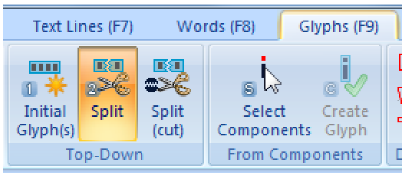


He Poſition: That Juſtified perſons
cannot totally nor finally, fall from the
grace of Union with the Lord Jeſus.

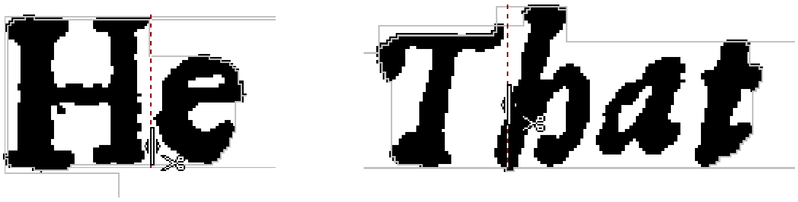
- Create glyph objects by either:
  - Splitting words
  - Marking single glyphs
    - Semi-automated or
    - Manually

3.1 Splitting words

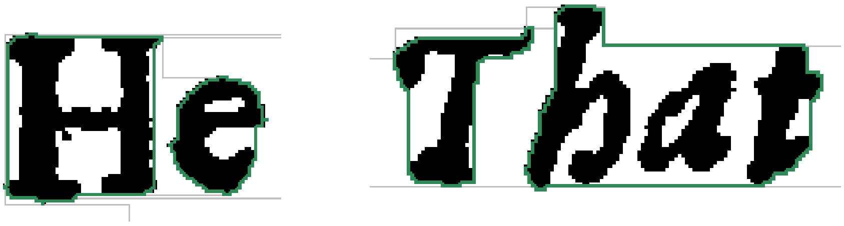
- Switch to the ‘Glyphs’ tab
- Activate the ‘Split’ tool (keyboard shortcut ‘2’)



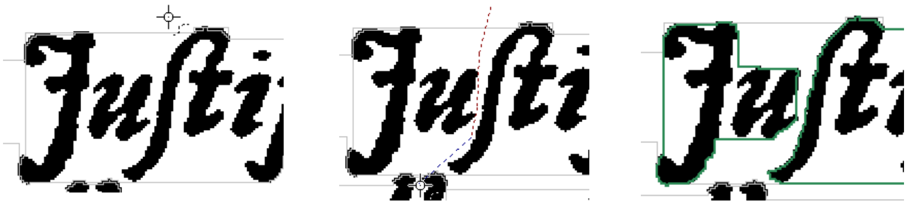
- Hover over the space between two characters of a word – a split line appears



- Click left to split



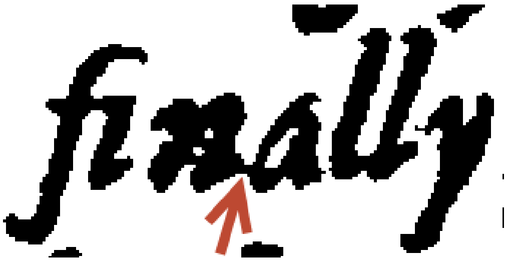
- For more complicated characters it is possible to define a split line with multiple segments:
    - Click left outside the word
    - Draw a split line between the two characters by continuously clicking left
    - Finish with right click or double click (latter adds a final point)
      Important: Finish the line outside the word.



Tip: Use the undo function if something went wrong.

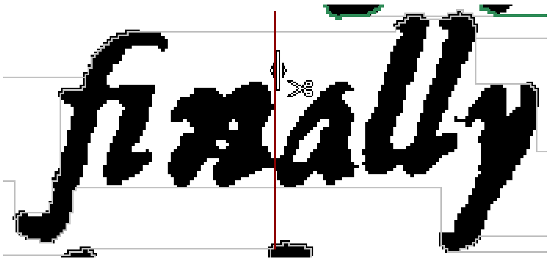
3.1.1 Special case: connected characters

In some cases, two or more characters can be connected within the black-and-white image:

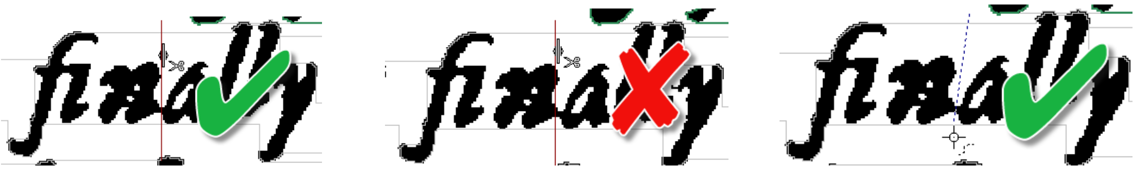


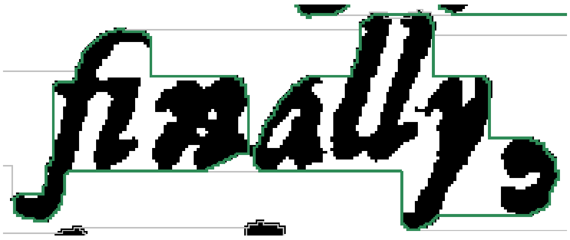
In this case proceed as follows:

- Activate the ‘Split (cut)’ tool (keyboard shortcut ‘3’)
- Hover over the space between two glyphs – a split line appears



- If the line does NOT cut through vital parts of the characters, click left to split, otherwise draw a multi-segmented split line as described above

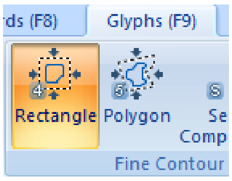




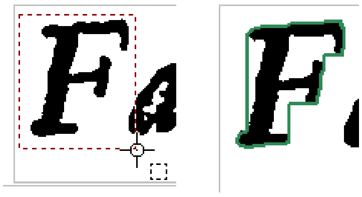
3.2 Semi-automated marking of glyphs

3.2.1 Contour detection starting from rectangle or polygon

- Switch to the 'Glyphs' tab
- Activate the 'Rectangle' or 'Polygon' tool from the 'Fine Contour' panel (keyboard shortcuts '4' and '5')

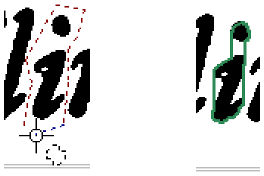


- Roughly mark the glyph by:
  - Drawing a rectangle (left click in one corner, drag, release in opposite corner)

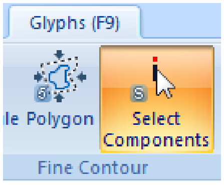
    OR

  - Drawing a polygon (left click for each polygon point, finish with right click or double click left (adds a final point))

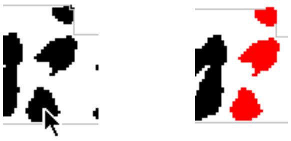
    

3.2.2 Contour detection for selected components

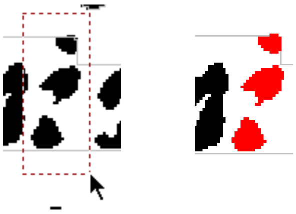
- Switch to the 'Glyphs' tab
- Activate the 'Select Components' tool (keyboard shortcut 'S')



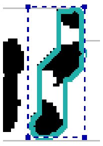
- Select all black components of a character by
  - Clicking left (use the CTRL key to add to current selection)



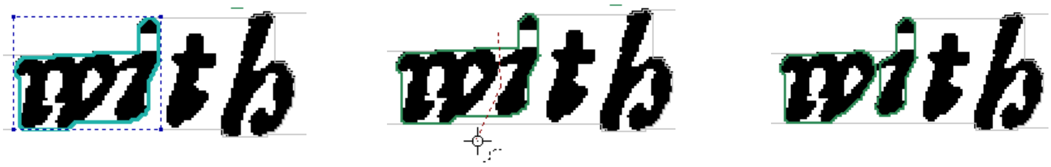
  - Or by dragging a selection frame (use the CTRL key to add to current selection)



- Click on ‘Create Glyph’ (keyboard shortcut ‘C’)

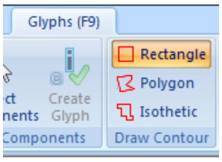


If two characters are connected in the black-and-white image, create one glyph object for both and split them as described in the previous section:

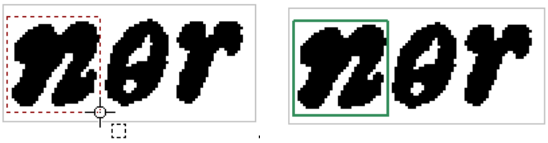


3.3 Manually marking glyphs

- To create a rectangular glyph:
  - Activate the ‘Rectangle’ tool (keyboard shortcut ‘R’)

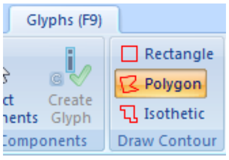


  - Draw a rectangle around a character by clicking and holding the left mouse button over one corner and dragging it to the opposite corner

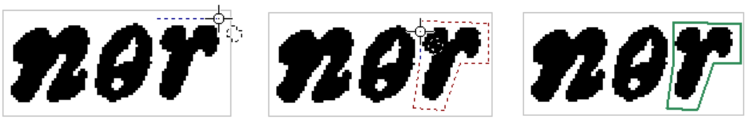


- Alternatively, click and release the left mouse button over one corner and right click over the opposite corner

- To create a polygonal glyph:
  - Activate the ‘Polygon’ tool (keyboard shortcut ‘P’)

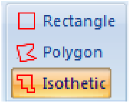


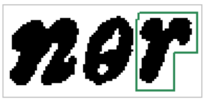
  - Draw a polygon around a character by continuously clicking left. Finish with right click or double click left (latter adds a final point to the polygon)



  - Alternatively, click and hold left and drag the mouse cursor around the glyph. Release the mouse when back at the starting point.

Hint: Isothetic polygons contain only horizontal and vertical line segments. They are usually more efficient for further processing.

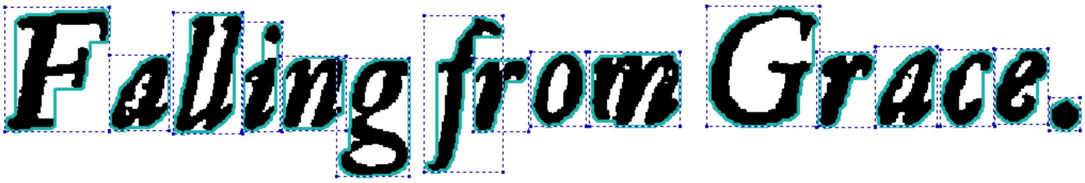




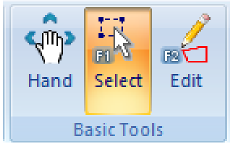
Important: Glyph objects should always be completely enclosed by their parent word object. Hence, to create glyphs manually, the word outlines should be spaced out sufficiently. Alternatively, word objects can be created after the glyphs (see chapter on bottom-up approach).

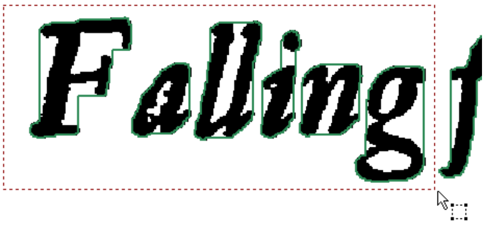
4 Bottom-up approach (from glyphs to regions)

- Create glyph objects as explained in the previous section (Semi-automated marking of glyphs and Manually marking glyphs)

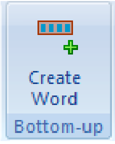


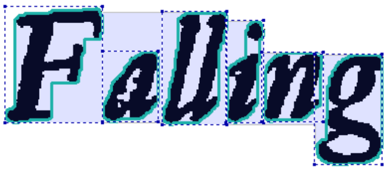
- Select all glyphs of a word by
  - Left click on each glyph (use the CTRL key to add to the current selection)
  - Or by using the Select tool which allows to drag a selection frame (keyboard shortcut F1)

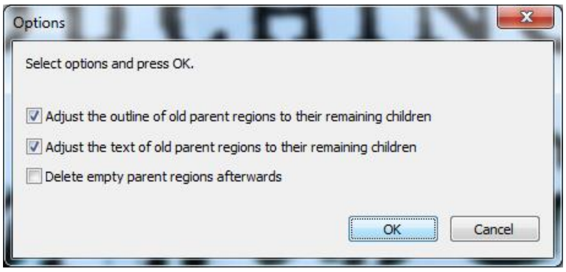




- Click on 'Create word' to create a word object for the selected glyphs (the dialogue with options is only relevant if the glyphs already belong to other word objects)







Options

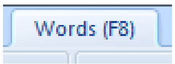
Select options and press OK.

☑ Adjust the outline of old parent regions to their remaining children

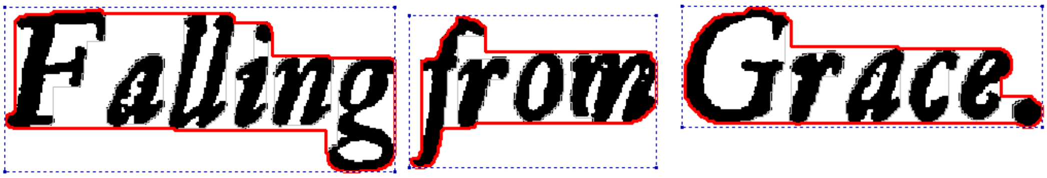
☑ Adjust the text of old parent regions to their remaining children

□ Delete empty parent regions afterwards

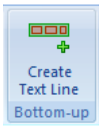
- Repeat for all words
- Switch to the 'Words' tab



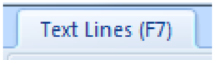
- Select all words of a text line



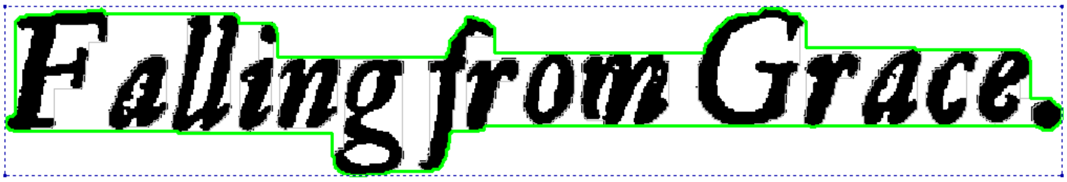
- Click on 'Create Text Line' to create a text line object for the selected words



- Switch to the ‘Text Lines’ tab

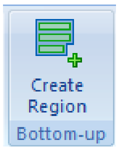


- Select all text lines of a region



Falling from Grace.

- Click on ‘Create Region’ to create a text region for the selected lines



5 Correcting pre-produced data

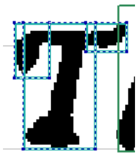
Prerequisites:

- Segmentation result in PAGE XML format
- Alternatively, the in-built Tesseract OCR engine can be used to segment a document page image

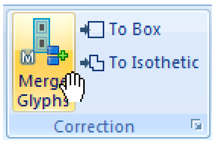
Note: Segmentation results should be corrected from top (regions) to bottom (glyphs). This tutorial only describes how to correct glyph objects.

5.1 Merging and splitting faulty glyph objects

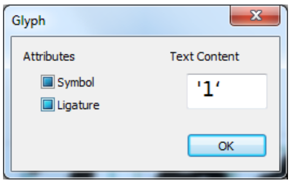
- To merge broken glyphs:
  - Select all parts by left click on each (use the CTRL key to add to the current selection)

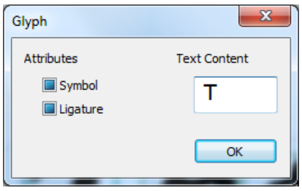


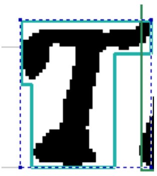
  - Click on ‘Merge Glyphs’ (keyboard shortcut ‘M’)



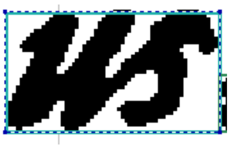
  - Correct the text content if necessary and click OK

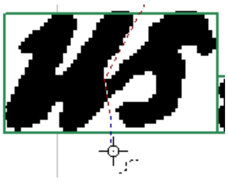


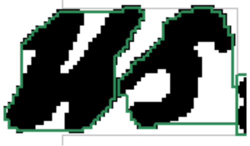




- To split connected glyphs follow the steps described earlier (Splitting words)



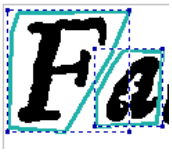




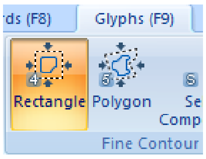
5.2 Contour Detection (‘Shrinking’)

The contour of an existing glyph object can be recalculated based on the black components that are within the glyph outline:

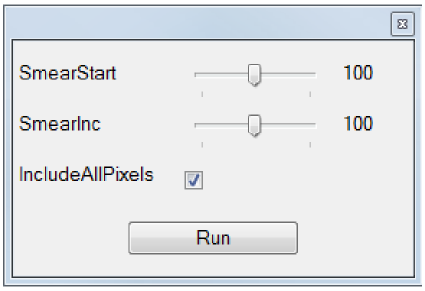
- Select one or multiple glyph objects



- Activate the ‘Rectangle’ or ‘Polygon’ tool from the ‘Fine Contour’ panel (keyboard shortcuts ‘4’ and ‘5’)



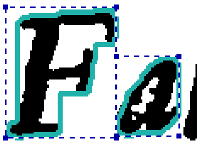
- Click on ‘Run’ within the tool window



SmearStart 100

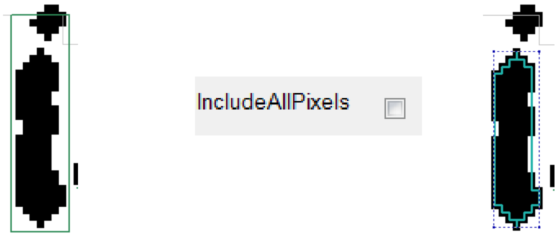
SmearInc 100

IncludeAllPixels ☑



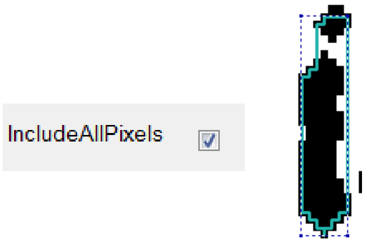
‘Include all pixels’ Option

By default (include all pixels disabled) the tool only takes into account black components that are completely inside the current object outline. Components that are partly outside are disregarded during the shrinking process. This might lead to unexpected results. See following example:



IncludeAllPixels ☐

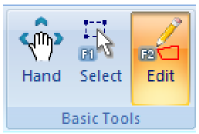
Tick the checkbox ‘IncludeAllPixels’ to include all black pixels that are within the current object outline:



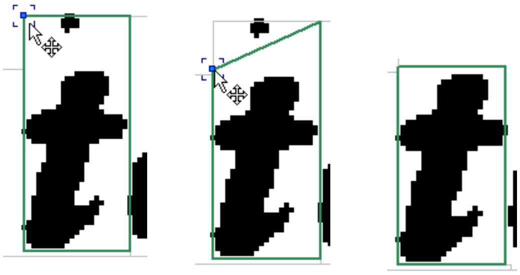
IncludeAllPixels ☑

5.3 Editing glyph outlines

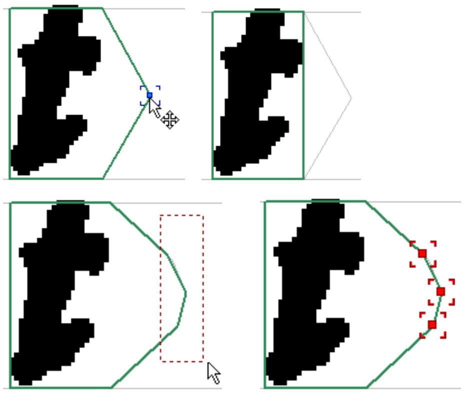
- To correct faulty polygonal outlines:
  - Activate the ‘Edit’ tool (keyboard shortcut F2)

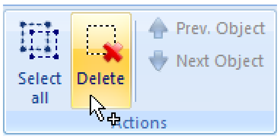


  - To move points, click and hold the left mouse button (Hint: use the CTRL key while moving a point to align it with its neighbour points)

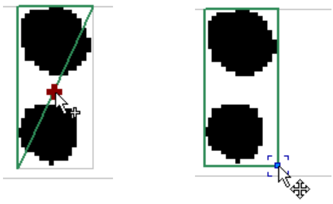


  - To delete points, hover over a point and press the delete key (alternatively select one or multiple points and click on ‘Delete’)





  - To add polygon points, hover over a polygon segments and click left (Hint: hold the mouse button to move the new point around)



Tip: Use the ESC key to quickly switch back to the ‘Hand’ tool from any other tool.

5.4 Adding missing glyphs

- To create new glyph objects follow the description from chapters Semi-automated marking of glyphs or Manually marking glyphs