Extraktion von Glyphen mit *Aletheia*

1. Zielsetzung

Den Anfang eines erfolgreichen OCR-Trainings bildet immer ein ausreichend großes und sauberes Glypheninventar, welches später die Trainingsgrundlage einer jeden OCR-Maschine bildet. Um nun das komplette Alphabet mit sämtlichen Sonderzeichen in einer ausreichenden Quantität zu erfassen bedienen wir uns eines von PRImA bereitgestellten Tools namens *Aletheia* zur Vorbereitung der Daten für das anschließende Training mit *Franken+.*

2. Funktionen von *Aletheia*

Aletheia ist das Flaggschiff einer Vielzahl von Tools, die *PRImA* dem interessierten Benutzer zur Verfügung stellt. Es ist ein System zur Analyse und Annotation digitaler Dokumente welches dem Nutzer die Möglichkeit bietet Regionen, Textzeilen, Wörter, Glyphen und viele weitere Informationen auszuzeichnen. Von diesem Funktionsumfang ist für unseren Workflow allerdings nur ein kleiner Teil interessant, nämlich die Funktion Glyphen zu kennzeichnen und diese anschließend mit ihren entsprechenden Unicodes auszuzeichnen. Das Ergebnis wird dann als PAGE-XML exportiert und ist damit bereit zur Weiterverarbeitung in *Franken+* für das Training von *Tesseract* importiert zu werden.

3. Aletheia 3.0 Pro vs. Lite

An dieser Stelle möchte ich kurz auf das Lizensierungsmodell und auf die wichtigsten Unterschiede der verfügbaren Versionen von *Aletheia* eingehen. *PRImA* bietet zwei Versionen von *Aletheia 3.0* an: Eine *Lite*-Version mit eingeschränkten Funktionsumfang und eine *Pro*-Version mit vollem Funktionsumfang inklusive einiger experimenteller Funktionen. Die *Lite*-Variante bietet keine Funktion um Bilder zu binarisieren, wohingegen die Pro-Version die drei am häufigsten genutzten Binarisierungsmethoden (Manueller Threshold, *Otsu*, *Sauvola*) sowie die Möglichkeit zur Bereinigung von Schmutz bietet. Ebenfalls fehlt der Lite-Version die Möglichkeit zur vollautomatischen Seitenanalyse sowie die Fähigkeit der Texterkennung. Die Lite-Variante beinhaltet grundsätzlich alle essentiellen Funktionen, die man benötigt, um Dokumente vollständig auszuzeichnen, allerdings sieht die Lizenz nur die Benutzung für persönliche Zwecke vor. Ein kommerzieller oder akademischer Einsatz ist mit dieser also nicht möglich. *PRImA* bietet für *Aletheia Pro* eine Probelizenz von 30 Tagen an, in der sämtliche Funktionen ausführlich getestet und evaluiert werden können.

4. Workflow-Übung *Aletheia 3.0 Lite*

Kommen wir nun zum praktischeren Teil des Vortrags. Zuerst können Sie sich noch entspannt zurücklehnen wenn ich Ihnen die grundsätzlichen Arbeitsschritte demonstriere; danach sind Sie eingeladen, diese auch selbst durchzuführen. Wir haben Ihnen hierfür einige Laptops mit den benötigten Programmen bereitgestellt. Gerne können Sie auch kleine Gruppen bilden und gemeinsam arbeiten.

Zuerst starten wir ein neues Dokument und öffnen hierzu die bereitgestellte TIFF-Datei. Dies ist eine Seite aus der 2ten Auflage des Narrenschiffs von Sebastian Brant, gedruckt 1494 von Peter Wagner in Nürnberg. Wir fahren fort ohne ein Farbbild zu wählen.

Wechseln wir auf den Reiter Glyphen haben wir nun alle benötigten Werkzeuge am oberen Bildschirmrand. Mit dem Tastenkürzel *S* oder durch den Knopf *Connected component selection* können wir jetzt einzelne Buchstaben selektieren und diese mit dem Tastenkürzel C oder dem Knopf *Create glyph from selection* in eine Glyphe verwandeln werden welcher den von Franken+ offiziell geforderten Polygon-Umriss besitzt.

Falls zwei oder mehr gewünschte Glyphen zusammenhängen, wandeln wir diese zunächst gemeinsam in einen Glyphe um; anschließend können wir mit dem Tastenkürzel 3 oder dem Knopf *Split (cut)* den Glyphen vertikal in zwei Einzelglyphen teilen.

Zerfallene Buchstaben sowie Buchstaben die standardmäßig aus mehreren Teilen bestehen, wie beispielsweise das i, müssen gemeinsam markiert und als eine Einheit in einen Glyphen umgewandelt werden. Um dies zu erreichen markieren wir zuerst einen Teil des Buchstabens wie gewohnt, anschließend fügen wir per Strg+Linksklick alle relevanten Komponenten hinzu. Sind alle Glyphenbestandteile markiert wird die Glyphe wie gewohnt erstellt.

Sind nun alle gewünschten Glyphen erstellt, müssen diese mit entsprechenden Text ausgezeichnet werden. Hilfreich ist hierbei vorher die Funktion *Text Overlay* zu aktivieren, welche den ausgezeichneten Text über den Glyphen erscheinen lässt. Markieren wir nun die erste Glyphe der Seite, indem wir vorher auf das Select-Tool per Testenkürzel F1 wechseln. Mit einem Klick auf *Text Content* oder mit dem Tastenkürzel F11 öffnet sich das Texteingabefenster. In diesem kann nun der entsprechende Buchstabe direkt per Tastatur eingegeben, aus der von *Aletheia* bereitgestellten Sammlung von Sonderzeichen ausgewählt, oder per Hexadezimal-Code (G = 0047) eingegeben werde. Zur nächsten Glyphe gelangen wir am einfachsten mit der Taste Bild-↓.

Ist der Text vollständig ausgezeichnet, ist es Zeit, das Dokument zu speichern. Der Dokumentname sollte hierbei dem Namen des Binärbildes entsprechen; dies erleichtert nicht nur die spätere Zuordnung, sondern ist auch Voraussetzung für den erfolgreichen Import in *Franken+.* Da Franken+ allerdings nur das ältere PAGE-XML Format von 2010 unterstützt, muss dieses unter *File → Export → Legacy PAGE XML 2010 (Aletheia 2.1)* geschehen.

5. Hands-On Teil

Jetzt haben Sie selbst die Möglichkeit, das gerade gelernte anzuwenden. Haben Sie keine Scheu Fragen zu stellen, wir beantworten diese gerne.

Phillip Beckenbauer

JMU Würzburg