Ausblick

Die folgenden Überlegungen sind noch Zukunftsmusik. Keine der betreffenden Methoden ist in der aktuellen Programmversion implementiert. Wenn die äusseren Umstände oder inhaltliche Hindernisse mich aber nicht davon abhalten, wird sich das Programm in die skizzierte Richtung bewegen.

- Selbstverständlich wird die Erkennungsqualität des OCR-Moduls verbessert werden. Dabei wird ein wahrscheinlichkeitsbasierter Ansatz stärker zum Tragen kommen. Ausserdem sollte die Anwendbarkeit neuronaler Netze, hier v.a. von anderen Formen als dem nur unter grossem Zeitaufwand trainierbaren BPN, genauer untersucht werden.
- Geplant ist auch - in einer sehr viel späteren Programmversion - die Erkennung weiterer südasiatischer Schriften und Sprachen. An erster Stelle steht aus persönlicher Affinität das Tamil, das zudem interessante Überschneidungen mit der Sanskrit-Lexikalik bietet.
- Das Programm sollte die inhaltliche Struktur gegebener Texte in einem gewissen Grad selbst erfassen und darstellen können. Mögliche Ansätze sind das *Data mining*, aber auch die schon erwähnte Benutzung semantischer Netze.
- Um den Kreis zu schliessen: Über das reine Wiederfinden von Wörtern und Wortkombinationen hinaus sollte sich das Programm zu einem wirklichen Hilfsmittel für die philologische Arbeit entwickeln, d.h. es sollte Texte und Textausschnitte vergleichen und räumlich/zeitlich einordnen können. Ein Beispiel, das mich schon länger beschäftigt, ist die Einarbeitung einer digitalisierten Verbreitungskarte der indischen Flora. Das Programm könnte dann z.B. auf Basis der in einem Text vorkommenden Pflanzennamen Schnittmengen der Verbreitungsgebiete erstellen und so einen ersten Hinweis auf den möglichen Entstehungsort des Textes liefern. Ein weiteres Beispiel ist die Entwicklung der Sanskrit-Lexik. Vielleicht ließe sich durch Auswertung grosser Textcorpora ein Trend hin zu bestimmten Wörtern feststellen und durch dieses vorgegebene Wissen ein neuer Text zeitlich einordnen. Ausserdem könnte man das Programm dazu benutzen, Passagen eines Textes fehlertolerant in einem anderen wiederzufinden und so die Abhängigkeit bestimmter Autoren untereinander neu zu definieren. Erste erfolgversprechende Ansätze zum Vergleich verschiedener Textpassagen arbeiten z.B. mit neuronalen Netzen, die mit „Satzvektoren“ aktiviert werden, oder vollziehen den Algorithmus von SOM-Netzen in einem dreidimensionalen Raum nach, wodurch die verschiedenen Schwerpunkte in der Lexik auch optisch verdeutlicht werden können. Allgemein formuliert: Das Programm sollte sich in künftigen Versionen von der Fixierung auf die reine Textgestalt und ihre Auflösung freimachen und sich stärker der inhaltlichen Analyse des Geschriebenen widmen.