Problemstellung und Einführung

Der Zweck des Programms, das hier vorgestellt wird, ist die weitgehende Automatisierung von Standardaufgaben im Bereich der Sanskrit-Philologie. Um diese Zielstellung zu erklären, sei ein Ausflug in den Beginn des Programms erlaubt.

Die ursprüngliche Zielsetzung dieser Dissertation war ein spezialisiertes Lexikon zur Erfassung von Pflanzennamen, Mineralien und komponierten Substanzen, die im Corpus des Ayurveda erwähnt werden, und der medizinischen Eigenschaften dieser Stoffe. Dazu war eine Datenbank entworfen worden, in der diese Informationen miteinander verknüpft werden konnten. Neben der Zuordnung von Eigenschaften zu Substanzen und der möglichst vollständigen Aufführung der Synonyme konnten auch Textvarianten gespeichert und einfache Handlungsabläufe - z.B. die Schritte, die zur Herstellung eines bestimmten Medikaments erforderlich sind - beschrieben werden.

Nachdem ein halbes Jahr vergangen und die Datenbank bei manueller Eingabe auf einige zehntausend Einträge angeschwollen war, andererseits aber gerade erst zwei relativ kurze Texte erfasst waren, kamen mir Zweifel an der Effizienz des Verfahrens. Ein weiteres Problem war die Modifizierung grundlegender Aufnahmekriterien, die einen mehrfachen Neubeginn bei der Codierung der Texte in der Datenbank erforderlich machte. Hier entstand der Gedanke, nicht manuell jeden Eintrag aus einem gedruckten Text zu extrahieren, sondern dem Programm einen digitalen Text zu präsentieren, aus dem es nach einfachen Kriterien Relationen zwischen Begriffen, z.B. Pflanzennamen und ihre Synonyme, in vorher markierten Textbereichen auslesen kann.

Nach anfänglichen Experimenten mit der manuellen Eingabe vollständiger Texte - eine ermüdende und nicht sehr belohnende Aufgabe - ging ich zu marktüblichen OCR-Programmen über. Hier zeigte sich schnell die Notwendigkeit, entweder ein Texterkennungsprogramm zu entwerfen, das die Besonderheiten des Devanagari-Alphabets erfassen kann, oder mit der manuellen Eingabe fortzufahren bzw. ein anderes Dissertationsthema zu wählen. Das vorliegende Programm ist das Ergebnis des ersten Weges.

In seiner aktuellen Version besteht das Programm aus drei Kernkomponenten:

- Ein OCR-Modul, das auf die Devanagari spezialisiert ist, ermöglicht es, gedruckt vorliegende Sanskrit-Texte über den Scanner direkt zu digitalisieren.
- Ein frei erweiterbares Lexikon auf der Basis des Sanskrit-Wörterbuchs von Monier-Williams speichert einen grossen Teil des relevanten Vokabulars der Sanskrit-Literatur.
- Ein Spracherkennungs-Modul kann digitalisierte Texte in ihrer lautlichen Originalform, d.h. ohne die sonst übliche Markierung von Sandhis usw., lexikalisch und grammatikalisch analysieren und die Analyseergebnisse dauerhaft in der Programm-Datenbank speichern. Wörter, die in einmal analysierten Texten auftreten, können damit jederzeit wiedergefunden werden.

Die folgenden Kapitel führen kurz in die Problematik dieser drei Bereiche ein und skizzieren dann die wichtigsten Algorithmen, die zur Lösung der Probleme entwickelt wurden. Ein abschliessender Abschnitt gibt eine Aussicht auf künftige Entwicklungen des Programms.

Ich danke Herrn Prof. Dr. Falk für seine tatkräftige Unterstützung dieser Arbeit. Aus der gemeinsamen Arbeit am IndoSkript-Projekt sind viele wertvolle Anregungen in das Programm eingeflossen.

Fr. H. Schott danke ich für die Gestaltung des Logos. Die grafische Gestaltung der Benutzeroberfläche erfolgt ausdrücklich ausserhalb ihres Verantwortungsbereichs.