Semantisch-Soziale Netzwerkanalyse am Beispiel buddhistischer Texte in der Pali-Sprache:

Zwischenstand zur Korpus-Aufbereitung

von Jürgen Knauth und Sven Wortmann

Projekt: SeNeReKo (Uni Bochum/ Uni Trier)

Ziel des Projekts „Semantisch-Soziale Netzwerkanalyse als Instrument zur Erforschung von Relgionskontakten“ (SeNeReKo) ist es maschinelle Analyseverfahren auf antike religiöse Textkorpora anzuwenden um herauszufinden, welche semantisch zentralen Begriffe mit welchen religiösen Akteuren verknüpft werden sowie die narrativen Muster von interreligiösen „othering“-Prozessen aufzuzeigen. SeNeReKo besteht aus zwei Teilprojekten zu altägyptischen Texten und buddhistischen Pali-Texten.

Im Teilprojekt zum buddhistischen Pali-Kanon werden große Textmengen in der mittelindischen Sprache Pali - einer indoeuropäischen Sprache ähnlich dem Latein und Altgriechischen - analysiert. Der buddhistische Pali-Kanon wurde etwa um die Zeitenwende auf Sri Lanka zusammengestellt, enthält die heiligen Texte des Theravada-Buddhismus (verbreitet in Sri Lanka, Thailand, Burma, Vietnam) und ist von großem Wert für die Rekonstruktion des religiösen Feldes des antiken Indien. In den Narrativen des Pali-Kanons finden sich unzählige Belehrungs- und Bekehrungsepisoden zwischen dem Buddha und buddhistischen Mönchen auf der einen Seite sowie religiösen Konkurrenten (brahmanische Priester und andere Asketen) auf der anderen Seite. Eine maschinelle semantisch-soziale Netzwerkanalyse dieser Texte wird genauer als herkömmliche manuelle Analysen die umkämpften Begriffsfelder und sozialen Konstellationen dieser Texte darstellen und somit den diskursiven Startpunkt einer der größten Religionen der Welt beleuchten. Ein Nebenertrag könnte darin bestehen, anhand von Wortfeld-, Sprachmuster- und Stilanalysen Spuren von innerkanonischem Texttransfer zu finden und somit die nach wie vor weitgehend ungeklärte Kompositionsgeschichte des Pali-Kanons zu erhellen.

Ausgehend von der Annahme, das für die von uns geplante Überführung der Textkorpora (bzw. von Teilen der Textkorpora) in semantische Netze die genaue Kenntnisse über die Wortart der einzelnen Wörter von großer Bedeutung sind, konzentriert sich der erste Teil der Arbeit in SeNeReKo auf eine geeignete Datenaufbereitung. Um eine Prozessierung der Texte zu einem späteren Zeitpunkt im Projekt zu erleichtern, wurde daher ein möglichst einheitliches Tagset festgelegt, welches nicht nur jeweils eine, sondern beide Textkorpora abdeckt. Aus Gründen der Erleichterung einer späteren Verwertbarkeit unserer Daten durch andere Wissenschaftler wurde bei der Erstellung dieses Tagsets auf Standards wert gelegt: Alle Tags verweisen daher auf Einträge in der ISOCat-Datenbank. Um die Texte mit geeigneten PoS-Tags auszuzeichnen, mussten jedoch textspezifische Wege beschritten werden, die sich im Altägyptischen und im Pali auf Grund der völlig unterschiedlichen Datenlage entsprechend voneinander unterscheiden. Das Ergebnis dieser Verarbeitung sind Texte im TEI-Format. Diese sind für Visualisierungen und weitere Verarbeitungsschritte über

einen Konverter in andere Datenformate überführt worden und liegen daher neben TEI ebenso in TCF-Format vor, sowie in einer HTML-Repräsentation.

Die Ausgangsdaten des Pali-Kanons waren eine Gruppe von etwa 2700 Einzeldateien in rudimentärer TEI-Auszeichnung vor. Der Strukturierungsgrad dieser Daten war nur wenig höher wie Fließtext. Daher wurde zu Beginn des Projektes eine Segmentierung und Tokenisierung der einzelnen Texte des Korpus Sätze, Wörter – und soweit vorkommend – sonstige Daten vorgenommen und in einem geeigneten TEI-Format gespeichert. Fast alle im Ausgangsmaterial vorhandenen Informationen wurden dabei beibehalten, so dass dieser Schritt im Grunde eine invertierbare Transformation darstellte.

In der weiteren Aufbereitung der Daten sollen die in den TEI-Daten enthaltenen Wörter mittels automatisiertem PoS-Tagging ausgezeichnet. Auf Grund des im Vergleich zu z.B. Latein geringeren Reichtums des Pali an Wortendungen ist ein statistisches Modell als Grundlage für ein PoS-Tagging unumgänglich. Daher wurde in Hinblick auf die Erzeugung eines solchen statistischen Modells 1000 Sätze nach dem Zufallsprinzip aus dem Korpus extrahiert, die manuell getaggt wurden (und weiterhin getaggt werden). Ein größerer Teil dieser Sätze liegt nun bereits in vollständig annotierter Form vor und kann für die wissenschaftliche Arbeit verwendet werden.

Der Auszeichnungsprozess mit Part-of-Speech-Tags ist im vorliegenden Fall einigen Besonderheiten unterworfen: Zum einen handelt es sich beim Pali um eine historische, tote Sprache. Die Auszeichnung muss daher durch nicht-Native-Speaker geschehen. Zum anderen finden sich in Pali Sandhis (Laufveränderungen), welche im vorliegenden Tagging-Prozess gesondert beachtet werden: Diese werden während des Taggens manuell aufgespalten. Geeignete Daten für die zukünftige Verarbeitung müssen Sandhis in bereits aufgelöster Form enthalten, daher ist bereits in unserem PoS-Tagging diese Aufspaltung vorgesehen. Ein Tagging-Werkzeug, welches für diesen Arbeitsschritt verwendet wird, müsste dies unterstützen. Da abgesehen davon neben einem besonderen Augenmerk auf gute Usability ferner Hilfswerkzeuge benötigt wurden und werden, die das Taggen von Teilen des Korpus vereinfachen, wurde die Entwicklung eines eigenen Tagging-Tools initiiert. Aus Gründen der Benutzerfreundlichkeit und Realisierungseffizienz basiert dieses Werkzeug nicht auf Web-Technologien, sondern ist als Desktop-Applikation angelegt. Es speichert jedoch analog zu Web-Anwendungen alle relevanten Daten auf einem Server und kann damit Anforderungen erfüllen, die in der Regel sonst nur von Web-Anwendungen erfüllt werden.

Das Tagging-Tool wurde auf Grund der begrenzten Projektkapazitäten auf die spezifischen Anforderungen im SeNeReKo-Projekt abgestimmt, ist aber grundsätzlich universell nutzbar angelegt. So beinhaltet es zwar auch projektspezifische Besonderheiten wie das Auflösen der oben erwähnten Sandhis, doch kann es mit auch in Zukunft für andere wissenschaftlichen Projekten genutzt werden. Vom effizienten User-Interface könnten gerade Projekte zu historischen Daten in Zukunft profitieren.

Die durch dieses Werkzeug erstellten Pali-Referenz-Daten stehen nun als Trainingskorpus für das Erstellen von Modellen für maschinelles PoS-Tagging zur Verfügung. Dieser Arbeitsschritt ist gegenwärtig „Work in Progress“.

Den Prozess des PoS-Taggings muss ein Prozess der Lemmatisierung des zu bearbeitenden Pali-Korpus begleiten, damit später semantische Netze korrekt erzeugt werden können: Wörter in flektierter Form sind in unserem Fall für eine Aufbereitung in solche Netze nicht nutzbar. Da eine solche Lemmatisierung wird ein umfassendes Datenbank aller Wortformen benötigt wird, wurde versucht, eine solche Datenbank zu realisieren. Diese kann prinzipiell aus den Einträgen eines regulären Pali-Dictionaries erzeugt werden.

Da es sich im Pali um keine gängige (lebende) Sprache handelt, gibt es leider jedoch kein vollständiges, computerlinguistisch nutzbares Wörterbuch des Pali. Daher konzentrierte sich ein Teil der bisherigen Arbeit auf die Aufbereitung eines bestehenden (gedruckten) Wörterbuchs der Pali Text Socienty, welches wir in rudimentärer digitaler Form von der Library of Chicago erhalten konnten. Die aufbereiteten Daten wurden dabei in einen eigens für das Projekt entwickelten Wörterbuchserver eingespeist, der dahingehend konzipiert und entwickelt wurde, dass unterschiedliche Personen und unterschiedliche Werkzeuge in Zukunft parallel und unabhängig voneinander mit ein und demselben Wörterbuch arbeiten können: Dem Aspekt der Verteiltheit und aktiven Arbeiten mit den Daten kommt hierbei besondere Bedeutung zu.

Die Realisierung dieses Wörterbuchservers erfolgte auf Grund eines möglichst schnellen Entwicklungszyklus und möglichst guter Performance durch Verwendung der NoSQL-Datenbank „MongoDB“ und einer davor gesetzten NodeJS-Webapplikation. Es handelt sich also dabei um eine klassische 2-Tier-Architektur. Hier treffen sich informatische Gesichtspunkte, insbesondere Aspekte der Softwarearchitektur mit Anforderungen der Digital Humanities: Durch diese Architektur wird nicht nur sicher gestellt, dass (wörterbuchserverspezifische) Applikationslogik nicht in unterschiedlichen Softwarewerkzeugen wiederholt implementiert werden muss, sondern gleichzeitig auch eine klar definierte (HTTP-basierte) Schnittstelle bereit gestellt, welche eine Unabhängigkeit darauf zugreifender Software von konkreten Programmiersprachen sicherstellt.

Der Server ist so gebaut, dass er Batch-Operationen unterstützt, um so netzwerkbedingte Latenzen bei den Zugriffen auf einzelne Wörterbucheinträge zu minimieren. Die Kombination dieses Konzepts mit den oben genannten Technologien erlaubt durchschnittliche Lese- und Schreiboperation von deutlich unter einer Millisekunde pro Wörterbucheintrag, so dass trotz des netzwerkbasierten Ansatzes und der dadurch normalerweise zu einem Problem auflaufenden Netzwerklatenzen es dennoch nicht zu performance-Problemen kommt, welche die Arbeit mit den Daten behindern könnten.

Zusammen mit einer an die Konzepte von Weblicht angelehnten Verwaltung einzelner softwaretechnischer Werkzeuge und einer XML-basierten Textdatenbank, die Textkorpus-Daten in TEI und TCF-Format liefern wird, steht in Kürze eine Infrastruktur zur Verfügung, welche eine saubere Datenverwaltung und Verfügbarhaltung für andere Projektteilnehmer gewährleistet. Gleichzeitig wird sie so den Aufwand für die nächsten Entwicklungsschritte reduzieret und es so ermöglichen, das die semantisch-soziale Netzwerkanalyse nicht nur in exemplarischen Beispielen oder fest vorgegebenen Teilen realisiert werden kann, sondern wahlfrei für beliebige Teile unserer historischen Textkorpora.