Titel: OpenSource-Bibliotheken und -Tools des SeNeReKo-Projekts

Autoren: Jürgen Knauth, Frederik Elwert

Abstract

Ziel des SeNeReKo-Projektes ist es, durch Techniken der **Se**mantisch-Sozialen **Ne**tzwerkanalyse einen Einblick in Teile von Textkorpora zu erhalten. Damit wird eine Form des „Distant Readings“ realisiert, im konkreten Fall zur Erforschung von **Re**ligions**ko**ntakten in altägyptischen Texten und dem Pali-Kannon. (= SeNeReKo)

Im Kontext des Projekts sind verschiedene Programmierbibliotheken und Werkzeuge entwickelt worden, um den Anforderungen des Projekts gerecht zu werden. Wesentliche Komponenten sind jedoch nicht projektspezifisch, sondern als allgemein verwendbare OpenSource-Komponenten geplant und umgesetzt worden: Wiederverwertbarkeit war von Vorneherein eines der Entwicklungsziele. Da Teilaufgaben von DH-Projekten durchaus öfters ähnlich gelagert sind, ist davon auszugehen, dass die so entstandenen Komponenten und Tools von anderen Wissenschaftlern entweder direkt oder nach geringer Adaption für andere Projekte genutzt werden können: Ziel des vorliegenden Posters ist es daher über genau diese Komponenten und Werkzeuge zu informieren. Da unsere Werkzeuge gerade deswegen entstanden sind, weil bislang noch nichts Vergleichbares zur Verfügung stand um die von uns angetroffenen Probleme effizient zu lösen, hoffen wir so durch unsere Software einen Beitrag für die Wissenschafts-Community zu leisten und so andere Wissenschaftler in ihrer zukünftigen Arbeit unterstützen zu können.

Konkret wurde in SeNeReKo ein Werkzeug zum Tagging von Texten entwickelt. Eine Besonderheit dieses Werkzeugs ist neben der Eigenheiten zur Auflösung von Pali-Sandhis seine besonders gut optimierte Usability: Unser Anliegen war hier, dass möglichst wenige Klicks erforderlich werden, um manuelle Tagging-Aufgaben durchzuführen. Das Fehlen von Tools mit vergleichbar Usability war Motivation der Entwicklung dieses Werkzeugs. Dieses client-server-basierte Standalone-Tool leistete einen wertvollen Beitrag in SeNeReKo für die Erstellung eines Gold-Standards im Pali, um weitere computerlinguistische Arbeitsschritte zu ermöglichen. Das Tool selbst kann jederzeit an die Verwendung für andere Sprachen angepasst werden.

Ferner stellen wir einen NoSQL-basierten Server zur Verwaltung von Wörterbuchdaten vor. Dieser ist als Komponente in einer klassischen Client-Server-Umgebung konzipiert und wird von den gleich nachfolgend erwähnten Werkzeugen verwendet: Ein Tool zur maschinellen Verarbeitung dieser Wörterbucheinträge, sowie einem Tool zur Visualisierung einzelner Datensätze. Der Server verwaltet dabei alle Wörterbucheinträge zentral und erlaubt dank seiner Bulk-Requests das effiziente Durchforsten der Daten auch bei größeren Datenmengen.

Ein Transformationswerkzeug, welches an den Server andockt, ist als IDE (Integrated Development Environment) konzipiert: Es erlaubt die Eingabe von C#-Programmcode-Fragmenten zur Datenverarbeitung. Diese Fragmente werden kompiliert; dann können sämtliche Wörterbucheinträge mit diesem Kompilat verarbeitet werden, um z.B. Muster zu erkennen und darauf basierend einzelne Wörterbucheinträge mit erkannten Informationen anzureichern. Eine Preview-Funktion gibt genauen Einblick darüber, auf welche Einträge sich die aktuell eingegebene Verarbeitungslogik erstreckt. Dadurch entsteht Transparenz: Erst der Einblick in die konkreten Änderungen über alle Datensätze hinweg erlaubt eine effiziente weil fehlerfreie Überarbeitung von Wörterbucheinträgen.

Ebenfalls an den Wörterbuchserver angegliedert ist ein Werkzeug zur Suche und Darstellung einzelner Wörterbuchartikel. Per serverseitig gespeicherter Konfiguration kann festgelegt werden, welche Controls in der GUI angezeigt werden sollen, und mit welchen Datenfeldern der einzelnen Artikel diese verbunden sein sollen: So kann eine Anpassung an Wörterbuchdaten beliebiger Struktur mit wenigen Handgriffen erfolgen. Die graphische Oberfläche erlaubt es, die manuelle Überarbeitungen auch größerer Artikelmengen auf einfachem Weg zu realisieren.

Ein anderes SeNeReKo-Tool unterstützt die Transformation beliebiger XML-Daten nach TEI: Über eine IDE-ähnliche Oberfläche können Umwandlungsregeln in Form eines Skripts eingegeben werden. Diese sind so gestaltet, dass sie fast schon natürlichsprachlich und somit leicht verständlich sind. Eine Erweiterbarkeit durch eigene Regeln ist jederzeit möglich: So können auch projektspezifische und über klassische X-Technologien möglicherweise nur schwer realisierbare Verarbeitungsprozesse durch ein Kommando repräsentiert werden (wie u.a. zur Verarbeitung im Pali in SeNeReKo). Angewandt auf eingelesene XML Datei(en) kann so die Aufbereitung von Daten erleichtert werden.

Des Weiteren stellen wir auf unserem Poster eine Reihe Tools vor, die dazu verwendet werden können, um aus TEI- bzw. TCF-Daten Netzwerke zu erzeugen. Sie stellen die Basis für den Kern des Projekts – die Erzeugung von Netzwerken dar. Da hier Standard-Graph-Datenformate für die Ausgabe verwendet werden, ist es möglich, diese Netzwerke dann anschließend mit verfügbaren Standard-Tools zu visualisieren.

Diese oben genannten Werkzeuge (bzw. Bibliotheken) sind frei nutzbar und helfen, gerade den teilweise sehr schwierigen Prozess der Datenaufbereitung zu adressieren. Wir würden uns freuen, wenn diese Komponenten nicht nur uns in SeNeReKo, sondern zukünftig auch anderen Wissenschaftlern helfen, damit so genau die Brücke geschlagen werden kann, die im Zentrum der Tätigkeit von uns Wissenschaftlern liegt: Die Brücke von Daten zu Erkenntnissen.