Protokoll 1 von 22.05.17

Ivana Daskalovska

Thema: *Optimierung der linguistischen Suche beim XML-annotierten Nachlass von Ludwig Wittgenstein*

Studentin: *Faridis Alberteris Azar*

Dozent: *Dr. Maximilian Hadersbeck*

Frau Alberteris Azar schreibt Ihre Bachelorarbeit im Rahmen des Digital-Humanities - Projekts „Wittgenstein in Co-Text“, im Zusammenarbeit mit dem Wittgenstein Archivs der Universität Bergen (WAB), in Norwegen. Für die Ziele dieses Projekts wurde am CIS-München eine Suchmaschine mit dem Namen WittFind entwickelt. Mit Hilfe dieser Maschine werden 5000 Seiten von dem Wittgensteins Nachlass (Typescript und Manuscript) bearbeitet.

Das Ziel der Bachelorarbeit ist die Optimierung der linguistischen Suche beim XML-annotierten Nachlass von Ludwig Wittgenstein durch die optimale Ausnutzung der XML-Annotation und die Verbesserung der XML-Edition mit Schwerpunkt auf Personennamen. Eine Erweiterung des Lexikons mit den gewonnenen Daten ist ebenso geplant.

Die Dateien aus den Bergen wurden an den Bedürfnissen von CIS in XML angepasst. Es existieren drei Typen von Dateien:

1. DIPLO – Der Inhalt wurde im Original gelassen
2. NORM – Der Inhalt wurde normalisiert
3. ORG – enthalten sowohl Original als auch normalisierte Form des Textes

Bei der Kategorisierung der Daten wird der TreeTagger (Probabilistischen POS-Tagger) verwendet, der von Herrn Dr. Helmut Schmid an der Universität in Stuttgart entwickelt wurde.

Viele Personennamen werden von dem Tagger nicht als Name Entitys erkannt. Es handelt sich überwiegend um Personennamen, die in geänderter, deklinierter Form im Text vorkommen. Mit der Hoffnung bessere Ergebnisse zu erzielen, haben die Wissenschaftler in den Bergen die Dateien aktualisiert und ein neues XML-Element hinzugefügt: PersNamen. Die neuen Dateien würden im März 2017 an CIS übergeben. Leider hat sich herausgestellt, dass diese Änderung keinen Erfolg gebracht hat, da der Tagger dieses Element bei der Zuordnung nicht berücksichtigt.

Frau Alberteris Azar versucht im Rahmen Ihrer Bachelorarbeit ebenso Möglichkeiten zu finden um bessere Ergebnisse zu erzielen.

In dem ersten Schritt versucht sie Eigennamen in NORM.xml Daten zu lokalisieren und Fehlerquellen zu sammeln. Sie nutzt dabei eine Schnittstelle von Python: The ElementTree XML API (etree bzw. ET)

In dem zweiten Schritt versucht sie die Semantische Suche in WittFind zu verbessern in dem sie Eigennamen in WittFind sucht, manche Einträge werden falsch als Eigenname zugeordnet.

Ihr Verbesserungsvoschlag hier wäre: neue syntaktische Kategorie „persName“ zu erzeugen:

1. in getaggte Dateien bei jeden gesammelten Beispielen „persName“ hinzufügen.
2. neue Kategorie in WittFind erzeugen:
   (([ADJA] | [ADJD])| [NN]) & & <persName>.
3. Neue Kategorie in CIS- Lexikon bei EN eintragen:
   Russellschen,Russell.EN+persName

Ergebnisse: Nach der Anwendung der neuen Prinzipien konnte man einen großen positiven Sprung merken. Mit Hilfe von WittFind wurden davor 168 Treffer in 13 von 20 Dateien gefunden, davon teilweise falsche Zuordnungen. Mit der Anwendung des verbesserten Systems wurden 833 richtige Treffer in allen Dateien gefunden.