Vortrag von Faridis Alberteris, Betreuer Dr. Maximilian Hadersbeck
Thema: Optimierung der linguistischen Suche beim XML- annotierten Nachlass von Ludwig Wittgenstein

Die Arbeit ist Teil des „Digital Humanities“ Projekts das in Zusammenarbeit mit der Universität Bergen (Norwegen) realisiert wird. Die Universität Bergen hat die Rechte an 20.000 Manuskripten des Wittgenstein Nachlasses. Ziel der Arbeit ist es die XML-Annotation der Daten optimal zu nutzen und die XML-Edition zu verbessern. Die XSLT-Dateien aus Bergen dienen zur Konvertierung der originalen XML-Editionen. Die Wittgenstein Daten sind in drei Gruppen unterteilt. Originaldateien (ORG), genormte Dateien und diplomatische Dateien. Die Originaldateien beinhalten alles, auch die Korrekturen die Wittgenstein selbst vorgenommen hat. Die genormten Dateien werden für Annotationsaufgaben verwendet. Die diplomatischen Dateien sind die korrigierten Originaldateien. Um Eigennamen in den genormten Dateien zu erkennen wird der Treetagger von Helmut Schmid verwendet. Der Treetagger basiert auf einem Markov-Modell. Der Tagger taggt die genormten Dateien und generiert genormte XML Dateien. Der Name „Tolstoi“ wird vom Tagger jedoch nicht erkannt, da der Name nicht in den Trainingsdaten auftaucht. Selbst wenn man die Trainingsdate aktualisiert, erkennt der Tagger den Namen immer noch nicht, da die aktuellen Informationen nicht genutzt werden. Um solche Fehler zu vermeiden werden 2 Schritte unternommen:

- Schritt 1: Eigennamen in Norm.xml Dateien lokalisieren
- Schritt 2: Semantische Suche in WittFind

In Schritt 1 werden zunächst alle möglichen Fehler beim Tagging gesammelt. Dazu verwendet man eine nützliche Schnittstelle in Python genannt: „The ElementTree XMLAPI“ (kurz: etree bzw. ET). Als Resultat werden Token Namen und Lemma Namen angezeigt. In Schritt 2 versucht man Eigennamen mithilfe von regulären Ausdrücken zu finden. Dazu verwendet man die syntaktische Kategorie „persName“, die zum regulären Ausdruck hinzugefügt wird. Durch diese Maßnahmen findet man jetzt in 13 Dateien 168 Personennamen.
Theme die in der Arbeit beleuchtet werden sind: Transkriptionsfehler vs. Editionsprobleme bei XML-Dateien im Wittgenstein Nachlass, Verbesserung der Tokenisierung, Verbesserung des Tagging und auf die bereits eingegangene Verbesserung der Personen Namenserkennung.
Eine weitere Fragestellung ist ob man das Lexikon erweitern kann, wenn man die Wortlist bzw. die Frequenzliste mit etree anstatt mit einem regulären Ausdruck erzeugt. Allgemein soll mit dieser Arbeit die linguistische Suche in WittFind verbessert werden, mit dem Schwerpunkt Eigennamenerkennung.