Protokol 1: Faridis Alberteris Azar „Optimierung der linguistischen Suche beim XML-annotierten Nachlass von Ludwig Wittgenstein“. Betreuer Hr. Dr. Maximilien Hadersbeck

Diese Bachelorarbeit ist auf dem Digital-Humanities Projekt „Wittgenstein in Co-Text“, im Zusammenarbeit mit dem Wittgenstein Archivs der Universität Bergen (WAB), in Norwegen basiert. Die Universität Bergen besitzt die Rechte für 10000 Seiten des „Traktatus“ vom Wittgenstein, 5000 davon XML transkribiert wurden CIS zur Verfügung gestellt.
Dafür hat CIS eine eigene Suchmaschine „Witfind“ entwickelt, die die Suche nach Wörter ermöglicht.

Das Ziel der Arbeit ist die Suche von XML-annotierten Nachlass von Ludwig Wittgenstein durch die optimale Ausnutzung der XML Annotation und die Verbesserung der XML-Edition zu optimieren. Das Schwerpunkt dabei liegt auf Personennamen.

Die XSLT Daten aus Bergen für die Konvertierung der originalen XML-Editionen haben drei Versionen: Originaldatei, die alle möglichen Schreibweisen von den Wörtern enthalten; Normalisierte Datei: enthalten nur die richtige Schreibweisen (diese Version wird für Witfind benutzt) und Diplomatische Datei enthält die Wörter genau so, wie Wittgenstein es geschrieben hat.

Für die Suche benutzt Witfind probabilistische POS-Tagger: TreeTagger, der von Herrn Dr. Helmut Schmid entwickelt wurde und MM basiert ist.

Tagger taggt die NORM- Dateien und generiert die NORM-tagged.xml. Beim Taggen wurden sehr viele Personennamen auf die alte Daten als solche nicht erkannt. Im März hat die Universität Bergen die aktuelle Version von Normalisierte Datei zugeschickt, wo die Personennamen extra einen Tag besitzen.

Die neue Datei enthält neue XML-Element: persNamen. Das Verhalten von dem Tagger auf die neue Daten hat sich im Bezug auf die negative Beispiele nicht positiv gewirkt.

In dieser Arbeit versucht man die XML Information nutzen, um die Suche zu optimieren.
Als erstes sammelt man einen Lexikon mit den Eigennamen. Weiterhin versucht man die semantische Suche anhand von den Regex zu verbessern. Man erzeugt eine neue syntaktische Kategorie.

Die Ergebnisse zeigen, dass für 20 Dateien findet das neue System 833 Treffer im Vergleich zum vorigen, wo nur 168 Treffer gab.

Im weiteren kann man sich mit den Transkriptionfehler und Editionsprobleme und Verbesserung der Tokenisierung beschäftigen.