Zusammenfassung zum Vortrag über die Arbeit „Optimierung der Linguistischen Suche beim XML-Annotierten Nachlass von Ludwig Wittgenstein“, gehalten von Faridis Alberteris
Zusammenfassung von Korbinian Schmidhuber

Ziel dieser Arbeit, die linguistische Suche in Wittfind zu verbessern, mit dem Schwerpunkt auf die Eigennamen und der Erweiterung des Lexikons. Wittfind ist ein am CIS entwickelte Suchmaschine auf dem Nachlass von Ludwig Wittgenstein.
Eigennamen führen in Wittfind oft zu fehlerhaften Ergebnissen, da diese von dem benutzten POS-Tagger nicht erkannt werden.

Es gibt verschiedene Dateien, die in dem Wittfind Projekt verwendet werden: Die Originaldateien, in denen alles mit aufgeführt ist, wie in den Manuskripten (mit Durchgestrichenem, Fehlern, usw.), diplomatische Version und eine normalisierte Version. Für diese Arbeti werden die normalisierten Daten verwendet. In diesen sind alle Wörter in der Form, wie sie richtig sind (Rechtschreibung, Groß-/Kleinschreibung).
Diese normalisierten Dateien werden in Wittfind z.B.. für den POS Tagger und die Erstellung der Frequenzlisten benutzt.

In Wittfind wird ein Probabilistischer Part-of-Speech Tagger benutzt. Dieser Tree Tagger basiert auf dem Markov Model, unterscheidet sich aber zu anderen POS-Taggern dadurch, dass er Entscheidungsbäume für das Messen von Übergangswahrscheinlichkeiten benutzt.

Für Wittfind gab es bis vor März diesen Jahres Norm-Dateien, die keine Namenstags hatten. Beim Taggen dieser Dateien kam es dadurch oft dazu, dass z.B. das Lemma „Tolstois“ (im Genitiv) nich als Name erkannt wurde. Dies hat zu teils fehlerhaften Ergebnissen geführt. Im März bekam das Institut neue Dateien von den Wittgenstein-Editoren in Bergen, die ein neues XML-Element besaßen: persNamen.

Im Rahmen dieser Arbeit wurde dieses XML-Tag in Wittfind eingebaut. Dabei wurde in folgenden Schritten vorgegangen: Zunächst wurden alle Lemma gesucht, die falsch getaggt wurden. Nützlich hierfür ist die ElementTree XML API in Python (etree oder ET).
Es wird pro Dokument eine Liste von Namen erstellt, in der Form Verwendung-Lemma, z.b. „Russel | Russel, Bertrand“. Im nächsten Schritt wird die semantische Suche in Wittfind verbessert, so dass Sie nun auch Eigennamen finden soll. Ein Beispiel, das bisher zu Fehlern geführt hat, war „hellenistisch“, das, ohne ein persNamen Tag als ein Name, wie in „Russel'sche“ interpretiert wurde. Eine solche Verbesserung soll dadurch erreicht werden, dass man eine neue syntaktische Kategorie, „persName“, in Wittfind erzeugt. Dabei sind 3 Schritte notwendig: diese Kategorie in den getaggten Daten hinzufügen, eine neue Kategorie in Wittfind erzeugen und eine neue Kategorie im CIS-Lexikon bei den EN eintragen.

Zuvor wurden 168 Personennamen gefunden, nach der Verbesserung waren es 833.

Weitere Ziele der Arbeit sind, Transkriptionsfehler finden und verbessern, die Verbesserung von Tokenisierung, des Tagging.
Zudem soll das Lexikon erweitert werden, dadurch dass die Wortlisten bzw. Frequenzlisten mithilfe von etree anstatt mit Regulären Ausdrücken erzeugt werden sollen.