

# Koreferenzauflösung

---

## Betreuer:

Peter Siniakov: [siniakov@inf.fu-berlin.de](mailto:siniakov@inf.fu-berlin.de)

## Zielgruppe:

Das Projekt richtet sich an Studenten im Hauptstudium, die Interesse an der Verarbeitung natürlicher Sprache haben. Linguistische Vorkenntnisse sind vom Vorteil, werden jedoch nicht zwingend vorausgesetzt. Kenntnisse der Grundtechniken der KI sind sehr wünschenswert.

## Motivation:

Koreferenzen sind Bezugnahmen auf dieselbe Entität an mind. 2 Textstellen. Menschen verwenden Pronomina als Koreferenzen gleichermaßen oft in gesprochener und geschriebener Sprache, um die Sprache mannigfaltiger, weniger eintönig, ästhetischer zu gestalten oder um die sprachliche Darstellung kompakter zu machen. Für automatische Verarbeitung stellen derartige Koreferenzen jedoch fast immer ein Problem dar, denn die Semantik eines Pronomens nur nach der erfolgreichen Koreferenzauflösung bestimmt werden kann. Diese Aufgabe wird oft durch den großen textlichen Abstand zwischen der Referenz und referenzierter Entität erschwert. Da aber Fakten oft mit Hilfe von solchen Koreferenzen ausgedrückt werden, ist eine zuverlässige Koreferenzauflösung für die Forschungsfelder Informationsextraktion und Textverständnis von großer Bedeutung.

Die Ergebnisse des Projekts werden in ein System zur Faktenextraktion aus natürlich-sprachlichen Texten einfließen, das im Rahmen des Forschungsprojekts FEx in der Arbeitsgruppe Datenbanken und Informationssysteme entwickelt wird.

## Ziele des Projekts:

Die in einem natürlich-sprachlichen Text vorkommenden Koreferenzen sollen gefunden und aufgelöst werden. D.h. die Verbindung zwischen der referenzierten Entität und jeder Referenz auf sie soll hergestellt werden.

## Durchführung:

Am Beginn des Projektes sollen die Teilnehmer sich in den aktuellen Forschungsstand vertiefen. Da es auf dem Gebiet der Koreferenzauflösung zahlreiche Ansätze gibt (das Spektrum reicht von wahrscheinlichkeitsbasierten Hidden Markov Modellen bis rein heuristischen Algorithmen), sollen die Ansätze kritisch betrachtet und hinsichtlich ihrer Tauglichkeit mit den Rahmenbedingungen des Projekts FEx bewertet werden. Linguistische Werkzeuge (morphologische Analyse, syntaktischer Parser) sind vorhanden und sollen eingebunden werden. Das entstehende Programm muss als eigenständiges Modul funktionieren, das einen Text mit semantischen Informationen über Koreferenzen in Form von XML-Annotationen anreichert.

Das Projekt soll von einer kleinen Gruppe durchgeführt werden. Start und Ende des Projektes können mit dem Betreuer flexibel vereinbart werden, genauso wie die regelmäßige Projekttreffen. Das Projekt beginnt mit einer kleinen Vortragsreihe, in der die

Projektteilnehmer kurze Vorträge über verschiedene Ansätze aus dem Bereich Koreferenzauflösung halten, um alle Projektteilnehmer auf den selben Wissensstand zu bringen. Danach erfolgt die Planung, Design und Implementierung des Tools.