Seit einiger Zeit werden im Informationsmanagement verstärkt Verfahren des Maschinellen Lernens eingesetzt, etwa in der automatischen Dokumentklassifikation, im Information Retrieval oder bei der Inhaltsaufbereitung von Dokumenten. Im vorliegenden Projektvorhaben wollen wir ausgewählte Verfahren des Maschinellen Lernens auf ihre linguistische Relevanz hin untersuchen.

Kann eine solche Relevanz, also ein wissenschaftlicher Erkenntnisgewinn, festgestellt werden, wäre es sinnvoll, diese Verfahren einzusetzen, um Corpusanalysen zu automatisieren und damit die Betrachtung größerer Datenmengen zu ermöglichen. Automatische Klassifikation und Clustering – beides Teilgebiete des Maschinellen Lernens – werden bereits erfolgreich für die Registeranalyse eingesetzt. Ein in der Corpuslinguistik noch wenig bekanntes Verfahren des Maschinellen Lernens sind die *Topic Models*. Topic Models sind probabilistische generative Modelle, die für sich in Anspruch nehmen, in einer Dokumentsammlung die thematischen Hauptstränge (die *Topics*) zu identifizieren, analog zu den Zielen der linguistische Analyse textbildender Merkmale (Kohäsion und Kohärenz), die sich allerdings auf einzelne Texte beziehen.

An der TU Darmstadt werden Verfahren zum Generieren von Topic Models wie etwa *Latent Dirichlet Allocation* (LDA) zum Beispiel zum Clustering von Produktmerkmalen in Corpora von Kundenbewertungen eingesetzt. Gegenstand des vorliegenden Projektvorhabens ist der Versuch, Analysen von Dokumentsammlungen mit Topic Models linguistisch zu interpretieren. Es ist stark zu vermuten, dass Topic Models sich nicht auf eine einzelne textbildende Kategorie abbilden lassen, sondern Aggregate verschiedener Kategorien sind (z.B. lexikalische Kohäsion, semantische Frames). Um mögliche Konvergenzen zwischen Topic Models und einer Analyse textbildender Merkmale feststellen zu können, soll an einem Corpus eine vergleichende Untersuchung durchgeführt werden. Ein Aufschluss über Gemeinsamkeiten/ Überlappungen zwischen Analysen auf der Basis von Topic Models und linguistischen Textanalysen ist aus Sicht der Linguistik von höchstem Interesse, kann sich hier doch eine weitere Möglichkeit der computertechnischen Unterstützung der sonst sehr aufwändigen, bisher nur manuell oder höchstens teilautomatisch durchzuführenden Textanalysen eröffnen. Zudem können sich nützliche Hinweise bzgl. einer Anwendung auf Textklassen (anstelle von einzelnen Texten)

ergeben, die ohne computertechnische Unterstützung nicht möglich wäre. Aber auch aus Sicht der
Informatik sollte eine solche Untersuchung von Interesse sein: Nur wenn stochastische Verfahren auch semiotisch interpretierbar sind, können sie überhaupt sinnvoll eingesetzt werden. Hier bieten linguistische Modelle einen möglichen extrinsischen Erklärungshorizont.