Abschlussbericht: Studentisches Forschungsprojekt LEHRE@LMU

Antragsteller:	Julian Baumann Studiengang Magister Englische Sprachwissenschaft, Phonetik und Physik
Projekt:	*An explorative approach to construction-based stylometry*
Forschungsziel:	Erforschung der Anwendung von *Konstruktionen* zur Korpus-gestützten stilometrischen Analyse von Texten (Magisterarbeit)

Einleitung *(modifiziert, dem Antrag entnommen)*: „Das ist wieder mal typisch BILD-Zeitung", „diese Geschichte ist ziemlich kafkaesk", „diese Doktorarbeit ist ein einziges Plagiat", „dieser Aufsatz wurde nie und nimmer von einem Viertklässler geschrieben!". Wir alle haben schon mal ein Wort, einen Satz, oder gar ganze Absätze oder Bücher gelesen und intuitiv vermuten können, wer der Urheber ist, oder wer es auf keinen Fall gewesen sein kann. Wie substantiell oder beweisbar dieses Empfinden ist, dass ein Text typisch oder untypisch für eine Person oder in Verbindung mit einer Sache sein kann, ist Aufgabengebiet der sogenannten *Stilometrie*. Im Laufe der Zeit haben sich unterschiedliche Auffassungen dieser Disziplin gebildet, die bis zum heutigen Tag trotz häufiger Anwendung keine anerkannten, standardisierten Analyseverfahren besitzt. Im Angesicht der steigenden Bedeutung in der Forensik und im Zivilrecht, entwickeln sich die zeitgenössischen, computer-gestützten Methoden der Stilometrie derzeit in eine Richtung, die aus linguistischer Perspektive fragwürdig erscheint, da die sprachliche Realität nur noch bedingt abgebildet wird. Die vorliegende sprachwissenschaftliche Studie verfolgt stattdessen das Ziel, Stilometrie mit Hilfe von *Konstruktionsgrammatik*, als einer innovativen Verknüpfung von traditioneller Ausrichtung und modernen, empirischen Methoden, zu erforschen.

Die erhaltenen Fördermittel wurden dabei für den Erwerb mehrerer Lizenzen einer dem wissenschaftlichen Standard entsprechenden Software für das so genannte POS-tagging (Wortartenklassifikation) von Texten eingesetzt. Diese war zum einen essentiell für alle empirischen Analysen im Verlauf der Studie, und wird zum anderen den Studierenden langfristig für eigene korpuslinguistische Projekte, sowie im Lehrbetrieb zur Verfügung stehen.

Ansatz: Bei dieser Studie handelt es sich um eine explorative Anwendung von quantitativen Methoden in der Linguistik. Die Umsetzung stützt sich dabei auf einen fundierten theoretischen Teil, in dem die Konzeptionen von „Stil", „Stilistischer Merkmale" und „Stilometrie" ausführlich erläutert und in Relation zu der Problematik moderner Analysemethoden und Sprache als Untersuchungsgegenstand gesetzt werden. Letzteres geschieht exemplarisch, indem repräsentative stilometrische Methoden aus zwei neueren Publikationen vorgestellt und kritisch evaluiert werden. Die Kritik besteht im Kern darin, dass die herangezogenen stilistischen Merkmale überwiegend als willkürlich isolierte (morphologische) Form-Einheiten verarbeitet werden, die ohne grammatikalischen und inhaltlichen Kontext die Kriterien und Funktion von Sprache nicht mehr erfüllen. Für das Ergebnis vieler quantitativer Analysen folgt daraus, dass, obwohl es möglich, ist Autoren mit einer hohen statistischen Verlässlichkeit voneinander zu unterscheiden, die quantitativen Unterschiede oft inhomogene Merkmale vereinen, die zudem nicht qualititativ ausgewertet können. Ein zentraler Ansatz der Studie besteht darin, dass die Quantifizierung von Sprache nicht ihre semiotische Natur (aus Zeichen/Form und Bedeutung) verletzen darf, sondern gerade diese Einheit als stilistische Information gezielt nutzen sollte. Dies ist nur möglich, indem stilometrische Analysen in einem fundamental anderen Framework für Sprache angewendet werden. Die Studie bedient sich daher der sogenannten Konstruktionsgrammatik, welche im Gegensatz zur stark restriktiven

„traditionellen Grammatik“, zahlreiche praktische Ansätze ermöglicht. Zu den wichtigsten zählen: (1) Die analysierten stilistischen Merkmale sind Konstruktionen, und umfassen als solche nicht nur einzelne Wörter oder Buchstaben, sondern lexikalisch-grammatikalische Einheiten, denen eine definierte Bedeutung zugrunde liegt. (2) Der unterschiedliche Gebrauch (die gemessene Häufigkeit) dieser Konstruktionen ist erklärbar über ein kognitives Modell, das besagt, dass die Verwendung von Sprache bei Menschen ein Ausdruck von unterschiedlich routinisiertem, und somit potentiell individuellem Verhalten ist. (3) Diese unterschiedlichen Routiniserungen können als „Präferenzen“ und „Abneigungen“ aufgefasst werden, wenn beispielsweise bei alternativen Ausdrucksweisen (Synonyme) statistisch signifikante Tendenzen erkennbar sind. (4) Gewisse Konstruktionen werden besonders häufig genutzt und sind potentiell individuell gebraucht, da sie Ausdruck grundlegender kognitiver oder emotionaler Erlebnisse sind, z.B. *I think that X* (‚Ich denke dass X‘), *I believe that* (‚Ich glaube dass X‘).

Durchführung: Ausgehend von diesen praktischen Implikationen wurde eine umfangreiche Analyse über 14 durch die Konstruktionsgrammatik gestützte Merkmale durchgeführt. Als Untersuchungsgegenstand dienten vier individuell zusammengestellte Korpora aus Werken der Autoren Arthur C. Doyle, James Joyce, Oscar Wilde und Virgina Woolf. Bei der Zusammenstellung und Vorbereitung der Korpora wurde darauf Wert gelegt, möglichst günstige Laborbedingungen für eine vergleichende Analyse zu schaffen. Die wesentliche Aufbereitung der Korpora wurde mit der erworbenen Software *CLAWS POS-tagger* (Part-of-Speech tagger) realisiert, entwickelt vom University Centre for Computer Corpus Research on Language (UCREL) der Universität Lancaster (UK). Für die Datenanalyse wurde das Konkordanz-Programm *AntConc* verwendet, entwickelt von Laurence Anthony an der Waseda University (JP). Die statistische Auswertung und Darstellung der Daten erfolgte in der Programmiersprache *R*.

Ergebnisse: Die empirischen Ergebnisse der Studie sprechen dafür, dass es über diesen Ansatz nicht nur möglich ist, stilistische Merkmale quantitativ zu erfassen, sondern auch einen bedeutenden qualitativen Mehrwert zu erhalten. Konkret stellt sich das Ergebnis nicht als statistischer Wert eines abstrakten Merkmals dar, sondern als statistische Auswertung vor dem interpretativen, semantisch-pragmatischen Hintergrund des betrachteten Merkmals. Diese Auswertung führen zu „Populationsmodellen“ aus denen signifikant individuelle „Präferenzen“ und „Abneigungen“ der Autoren gegenüber Ausdrucksweisen ersichtlich sind. Ob diese individuellen Tendenzen stark genug sind, um in einem Einzelvergleich Autor A signifikant von Autor B, C, oder D zu unterscheiden, wurde für jedes Merkmal in einer Serie von sechs statistischen Tests ermittelt. Die Merkmale selbst können dabei jederzeit auf ihren dementsprechenden semantischen Kontext zurückgeführt werden. So war es beispielsweise möglich, in einem Feld von quasi-synonymen Ausdrücken (*know, believe, think, feel that X*), individuelle Profile in einer semantischen Tafel zu illustrieren, indem sie den Dimensionen *high* vs. *low confidence* und *rational* vs. *emotional* zugeteilt wurden. Weitere Methoden basierten auf der statistischen Auswertung von assoziierten grammatikalischen Mustern, oder erfolgten auf der Basis linearer Korrelationskoeffizienten, welche über die semantische Rollenverteilung von *Agent* und *Rezipient* auf männliche und weibliche Pronomen Aufschluss geben konnten.

Weiterführung: Im Rahmen der Analyse konnte gezeigt werden, dass der gefundene Ansatz das Potential birgt, wertvolle Impulse zur Quantifizierung von Sprache zu liefern. Eine Fortführung der Studie wäre denkbar unter der Reduzierung von untersuchten Merkmalen bei einer größeren Anzahl von Autoren. Die bereits stark interdisziplinäre Ausrichtung dieser Studie könnte weitere Erforschung in einer fachübergreifenden Kooperation mit der Computerlinguistik oder statistischen Mathematik ermöglichen. Desweiteren wurden alle notwendigen Schritte zur Dokumentation und Bereitstellung der verwendeten Software für weitere Projekte Studierender bereits unternommen.