

# Programm

## Workshop Forensische Linguistik /Forensic Linguistics

Ruhr-Universität Bochum. Internationales Begegnungszentrum (IBZ) in Beckmanns Hof

17. November 2017

### Donnerstag, 16. November 2017

Ab 18:30 Warm-up beim Abendessen im [Livingroom](#)

### Freitag 17. November 2017

9.00 -9.30 Begrüßung durch Prof. Christof Paar, Sprecher des Graduiertenkollegs „Sicherheit für Menschen im Cyberspace“ (SecHuman)

9.45-10.30 Martin Potthast (Universität Leipzig): Tackling Societal Challenges with Style Analysis

10.30-11.00 Kaffeepause

11.00-11.45 Sabine Ehrhardt (Bundeskriminalamt): Der linguistische Textvergleich als kriminaltechnische Methode zur Auswertung geschriebener Sprache

11.45-12.30 Oren Halvani (Fraunhofer SIT Darmstadt): Autorschaftsverifikation auf der Basis von Kompressionsmodellen

12.30-14.00 Mittagspause

14.00-14.45 Karoline Marko (Universität Graz): Verstellungsstrategien in schriftlichen Drohungen und Erpresserschreiben

14.45-15.30 Tatjana Scheffler (Universität Potsdam): Linguistic Variability on Social Media - Constructing and Analysing Corpora

15.30-16.00 Kaffeepause

16.00-16.30 Abschlussdiskussion

Danach gibt es noch eine Führung durch Bochum für Interessierte, die noch nicht abreisen.

# Abstracts

## **Der linguistische Textvergleich als kriminaltechnische Methode zur Auswertung geschriebener Sprache**

Sabine Ehrhardt (BKA)

Der linguistische Textvergleich ist eine kriminaltechnische Methode zur Auswertung geschriebener Sprache, die für Ermittlungen und Gerichtsverfahren gleichermaßen bedeutsam ist und große Auswirkungen haben kann. Aus sprachwissenschaftlicher Sicht steht ein umfangreiches Instrumentarium zur Beschreibung von Texten zur Verfügung, auf das in vielen Publikationen Bezug genommen wird. Deutlicher seltener stehen jedoch der Aufbau eines Textvergleichs im kriminaltechnischen Kontext und die Auswertung sprachwissenschaftlicher Daten im Rahmen der Befundbewertung im Fokus des Interesses.

Der hier angekündigte Vortrag wird die sprachwissenschaftliche Arbeit in der Kriminaltechnik am Beispiel des linguistischen Textvergleichs darstellen und dabei insbesondere auf die Erfordernisse des forensischen Kontexts eingehen.

## **Autorschaftsverifikation auf der Basis von Kompressionsmodellen**

Oren Halvani (Fraunhofer-Institut Darmstadt)

Der Vortrag beschreibt zunächst das Forschungsfeld der Autorschaftsverifikation (AV) und verdeutlicht die Unterschiede gegenüber der verwandten Disziplin der Autorschaftsattributions. Anschließend wird eine einfache und effektive AV-Methode präsentiert, welche auf sogenannten Compression Models basiert. Dabei wird auch auf die Herausforderungen eingegangen, die hierbei auftreten. Der Vortrag schließt mit einem Überblick über die offenen Fragestellungen auf dem Gebiet der AV.

## **Verstellungsstrategien in schriftlichen Drohungen und Erpresserschreiben**

Karoline Marko, Universität Graz

Das Ziel dieser Untersuchung ist es, sprachliche Verstellungsstrategien in schriftlichen Drohungen und Erpresserbrieffen zu untersuchen. Dafür wurden zwei empirische Studien durchgeführt. Die 124 TeilnehmerInnen (68 ÖsterreicherInnen, 56 US-AmerikanerInnen) gaben persönliche Informationen zu ihrem demographischen und sozialen Hintergrund bekannt und mussten zwei Texte verfassen: ein informelles E-Mail und eine Drohung oder einen Erpresserbrief. Im ersten Teil der Analyse wurden sprachliche mit sozialen Variablen wie Geschlecht, Alter, Bildung und regionaler Herkunft korreliert. Die Ergebnisse der Analysen beider Textsorten wurden verglichen um herauszufinden, welche Unterschiede es zwischen dem informellen E-Mail und dem Drohschreiben gibt und welche sprachlichen Variablen von den Autoren wie manipuliert wurden. Etwa 20% der TeilnehmerInnen gaben an, Verstellungsstrategien verwendet zu haben. Eine Analyse zeigte aber, dass in etwa 40% der Fälle Hinweise auf eine Verstellung vorlagen. Die häufigsten Strategien waren: willkürliche Verstellung, Anonymisierung und Änderung des sprachlichen Registers.

Die zweite Studie mit 167 TeilnehmerInnen (58 US-AmerikanerInnen; 109 ÖsterreicherInnen) wurde durchgeführt, um herauszufinden, ob sich die Ergebnisse der ersten Studie bestätigen oder widerlegen lassen. Die TeilnehmerInnen wurden zu ihren Erfahrungen mit Drohschreiben, ihren Vorstellungen bezüglich Drohschreiben, Verstellungsstrategien und sprachlichem Profiling befragt. Der letzte Teil der Arbeit befasst sich mit der Anwendung der Ergebnisse auf authentische Texte (einen englischen Erpresserbrief und ein deutsches Drohschreiben). Es konnte gezeigt werden, dass es zu einem gewissen Grad möglich ist, die Ergebnisse von kontrollierten Studien auf authentische Texte anzuwenden.

## **Tackling Societal Challenges with Style Analysis**

Martin Potthast (Universität Leipzig)

The talk will give an overview of the work that the Webis group has done in the past years on employing stylometry to tackle some of the societal challenges of society today. The talk briefly describes recent work on clickbait detection, hyperpartisan news detection, fake news detection, and author obfuscation. We will also briefly review the PAN initiative as one of our main contributions to foster progress in research and practice with regard to digital text forensics.

## **Linguistic Variability on Social Media - Constructing and Analysing Corpora**

Tatjana Scheffler, Universität Potsdam

Social media such as Twitter, Facebook, blogs, forums, etc. have become huge collections of user-generated data online. Since many more people participate on these channels than in traditional media, this provides unique opportunities for studying linguistic variation. At the same time, the large amounts of data and the informal types of content pose specific challenges for researchers trying to collect and analyze social media texts. In this talk, we present computational linguistic methods for collecting specialized social media corpora and for analyzing the resulting data sets linguistically.

In the presentation, we show how we constructed a corpus of German social media data from the same users in Twitter and blogs, allowing us to study linguistic variability within the same individual and across media channels. Based on comparisons of social media corpora with traditional texts and spoken language, we have identified several phenomena in which social media conversations differ from texts: For example, they contain questions, particles, fill words, informal language and alternative spellings. These non-standard features of social media text are challenging for automatic processing, and we discuss approaches for either normalizing the data or adapting the tools to the specific properties of the data.