Workshop „Digitale Korpuserstellung und -auswertung“
(Nr. im VL: 04063790)
Freitag, 9.11. und Samstag, 10.11.2018
Uni Würzburg, ZHSG, Am Hubland
Organisiert und finanziert durch: CLiGS-Gruppe

		A: Sprachwissenschaft Raum 1.004	**B: Literaturwissenschaft Raum 1.003**
Freitag, 9.11.	14-15.30	Session I: *Einführung: Corpus Explorer* Jan Oliver Rüdiger (Kassel)	Session I: *Korpuserstellung: Manuelle Texterfassung und Daten-kuration* Matthias Boenig (Berlin, BBAW/OCR-D) & Benjamin Fiechter (Berlin, BBAW/ CLARIN-D)
	15.30-16	Kaffeepause	
	16-17.30	Session II: *Einführung: Corpus Explorer* Jan Oliver Rüdiger (Kassel)	Session II: *Korpuserstellung: Automatische Texterfassung* Matthias Boenig (Berlin, BBAW/OCR-D) & Benjamin Fiechter (Berlin, BBAW/ CLARIN-D)
	18-20	Abendvortrag (Raum 1.003): *Diskurstraditionen und Textkomplexität* Angela Schrott (Kassel)	
	ab 20.30	gemeinsames Abendessen	
Samstag, 10.11.	9-10.30	Session III: *Einführung in die Datenexploration und Visualisierung mit R* Malte Rosemeyer (Freiburg/Leuven)	Session III: *Digitale Literaturwissenschaften: Methoden im Überblick* Christof Schöch (Trier)
	10.30-11	Kaffeepause	
	11-12.30	Session IV: *Einführung in die Datenexploration und Visualisierung mit R* Malte Rosemeyer (Freiburg/Leuven)	Session IV: Einführung in die Textanalyse mit Regulären Ausdrücken und TXM Christof Schöch (Trier)
	ab 13	gemeinsames Mittagessen	

Inhaltliche Beschreibung

Sprachwissenschaft:

Einführung: CorpusExplorer (Jan Oliver Rüdiger)
Ziel des CorpusExplorers ist es, korpuslinguistische Methoden möglichst einfach und intuitiv zu bündeln. Das Programm verfügt über 50 unterschiedliche Analysen/Visualisierungen unter einer einheitlichen und leicht zu bedienenden Programmoberfläche. Damit richtet sich der CorpusExplorer nicht nur an Forschende sondern auch an Lehrende und Studierende. Neben typischen Quellen der Korpuslinguistik, wie digitale Zeitungstexte, oder Webseiten lassen sich auch E-Mails, Tweets und Transkriptdaten (z. B. CLAN/CHILDES, EXMERaLDA) einlesen und verarbeiten. Der oft sehr aufwändige Aufbereitungsprozess − Bereinigung, Metadatenextraktion, Annotation (z. B. mit Lemma und Wortarteninformationen) − wird durch das Programm vollständig automatisiert. Ziel des Workshops ist es: (1) Eine praxisnahe Einführung in den CorpusExplorer zu geben. (2) Diese Grundlage soweit zu verfestigen, dass erste Ideen zum eigenen Seminareinsatz entwickelt werden können. (3) Vertiefend - Interaktionsmöglichen zwischen R und dem CorpusExplorer aufzuzeigen.

Einführung in die Datenexploration und Visualisierung mit R (Malte Rosemeyer)
Die Open-Source Software R ist in den letzten Jahren zum Standardwerkzeug in vielen sprachwissenschaftlichen Disziplinen, besonders aber korpusbasierten und komputationalen Methoden geworden (vgl. z.B. Levshina 2015: 21). Dies liegt einerseits an ihrer großen Flexibilität und der Tatsache, dass R über eine Fülle an spezialisierten Funktionen und Arbeitspaketen für sehr spezifische Aufgabe verfügt. Andererseits wird immer wieder betont, dass R sich besonders gut zur Visualisierung von Ergebnissen eignet (vgl. z.B. http://r-statistics.co/Top50-Ggplot2-Visualizations-MasterList-R-Code.html <5.9.2018>).
Dieser Workshop soll eine kurze Einführung in die Exploration und Visualisierung von Daten anhand von R leisten. Wir werden im ersten Teil des Workshops grundlegende Funktionen in R zur Erzeugung von Tabellen und Berechnung von einfachen statistischen Maßen kennenlernen. Im zweiten Teil werden wir uns mit *ggplot2*, einem besonders hilfreichen Paket von Funktionen zur Erzeugung von Graphen in R beschäftigen, seine Syntax kennenlernen und selbst einfache Graphen erzeugen.

Literaturwissenschaft

Korpuserstellung: Manuelle Texterfassung und Datenkuration & Automatische Texterfassung
(Matthias Boenig & Benjamin Fiechter)
Der Workshop richtet sich an alle Forschenden, die für ihre Arbeit oder Sammlung eigenständig oder kollaborativ ein Textkorpus erstellen möchten. Ziel des zweiteiligen Workshops ist die grundlegende Vermittlung von Arbeitsmethoden und Organisationsformen (d.h. Verfahren und Prozesse, auch der gemeinsamen, arbeitsteiligen Korpuserstellung), die notwendig sind, um Textkorpora für die wissenschaftliche Forschung zu erstellen. Zudem werden Techniken und Verfahren der Datenkuration (bzw. ‘Digital Curation’) vermittelt, durch die bestehende Textressourcen zu interoperablen und nachnutzbaren Textressourcen aufgewertet und in bestehende Korpora integriert werden können. Nicht die verschiedenen Ansätze editorischer Schulen sollen im Mittelpunkt des Workshops stehen, sondern möglichst generische Richtlinien, Standards, Verfahren und Formate zur manuellen und zur automatischen Texterfassung vermittelt werden. Anhand von Beispielen wird jeweils gezeigt, wie der Prozess der Texterfassung aus verschiedenen Publikationsformen ((handschriftliche) Briefe, Tagebücher, Archivalien, Druckpublikationen (Zeitungen, Bücher), bestehende Datensammlungen)

erfolgen sollte. Der Workshop baut auf den Erfahrungen und Workflows der Projekte *Deutsches Textarchiv* (http://www.deutschestextarchiv.de/) und *OCR-D* (http://www.ocr-d.de/) sowie des Infrastrukturprojekts *CLARIN-D* (https://clarin-d.net) auf.

Digitale Literaturwissenschaften: Methoden im Überblick (Christof Schöch)
Diese Sitzung hat zum Ziel, einen Überlick über Methoden der digitalen Literaturwissenschaften zu vermitteln. Dabei wird ein besonderer Fokus auf quantitativen Verfahren liegen, darunter die stilometrische Autorschaftsattribution und thematische Exploration mit Topic Models und generell der Einsatz von maschinellem Lernen für die Bearbeitung literaturwissenschaftlicher Fragestellungen. Die Sitzung wird mit einer Diskussion der Möglichkeiten und Grenzen der quantitativen Literaturwissenschaften schließen.

Einführung in die Textanalyse mit Regulären Ausdrücken und TXM (Christof Schöch)
Diese Sitzung hat zum Ziel, einen praktischen Einstieg in die computergestützte, textnahe Analyse von kleineren bis mittleren Textsammlungen zu bieten. Eine wichtige Grundlage hierfür sind sogenannte Reguläre Ausdrücke, mit denen komplexe Muster von Zeichenketten in Texten gesucht werden können. Darauf aufbauend wird dann in das Textanalysetool TXM eingeführt, das weiterführende Such- und Analyseverfahren ermöglicht. Die Teilnehmer/innen an dieser praxisorientierten Sitzung sollten ihren eigenen Laptop zur Sitzung mitbringen.