Proposal für einen Pre-Conference Workshop zur
2. Jahrestagung *Digital Humanities im deutschsprachigen Raum (DHd)*

Computerlinguistische Methoden der Inhaltsanalyse in den Sozialwissenschaften: Forschungspraktische Herausforderungen, Werkzeuge und Technologien

Organisiert von den Kooperationspartnern des BMBF-Verbundprojekts *e-Identity*

Überblick

Zeit (Vorschlag):	Montag, 23. Februar 2015, 14:00 bis 19:00 Uhr Dienstag, 24. Februar 2015, 09:00 bis 13:00 Uhr
Organisatoren:	Prof. Dr. Manfred Stede, Jonathan Sonntag (beide Universität Potsdam), Prof. Dr. Cathleen Kantner, Maximilian Overbeck (beide IfS Stuttgart), Prof. Dr. Jonas Kuhn, Dr. André Blessing (beide IMS Stuttgart), Prof. Dr. Ulrich Heid, Fritz Kliche (beide Universität Hildesheim)
Teilnehmerzahl:	ca. 15 – 20
Technische Ausstattung:	Beamer (VGA)
Kontakt:	Projekt-Webpage: http://www.uni-stuttgart.de/soz/ib/forschung/Forschungsprojekte/eIdentity.html

Inhalt des Workshops

Aufgrund der dramatisch angestiegenen Verfügbarkeit großer Korpora sozialwissenschaftlich relevanter Textdaten erlebt die Forschungslandschaft der Sozialwissenschaften aktuell einen regelrechten Boom der Methoden für die computerlinguistische Inhaltsanalyse. Dabei werden die Akzente mal stärker auf quantitative Auswertungen von Texten, mal stärker auf qualitative Interpretation und Annotation von Textdaten gesetzt. Unser Workshop möchte die Perspektiven beider Seiten zusammenbringen und dabei insbesondere auch die Möglichkeiten der sinnvollen Ergänzung quantitativer und qualitativer Analyse in den Blick nehmen.

Die Organisatoren des Workshops sind in dem interdisziplinären Forschungsprojekt *e-Identity* vernetzt und untersuchen aus politikwissenschaftlicher und computerlinguistischer Perspektive die internationale Diskussion über Kriege und humanitäre Interventionen seit dem Ende des Kalten Krieges. Dabei stehen folgende Fragestellungen im Vordergrund: Wie mobilisieren internationale Akteure in Krisensituationen kollektive Identitäten? Spielen sie ethnische, religiöse, nationale, europäische, u.a. Bindungen gegeneinander aus? Welche Ursachen und Effekte hat diese Identitätspolitik? Das Projekt untersucht internationale Diskussionen über Kriege und humanitäre Interventionen seit dem Ende des Kalten Krieges. Das Forscherteam greift auf ein bereinigtes mehrsprachiges Korpus von mehreren hunderttausend Zeitungsartikeln aus der Qualitätstagespresse mehrerer europäischer Länder (Österreich, Deutschland, Irland, Frankreich, Vereinigtes Königreich) und den USA zurück (kontinuierlich erhobener Untersuchungszeitraum: Januar 1990 - Dezember 2011).

Um die Analyse dieser komplexen, theoretischen Konzepte auf großen Textkorpora von Zeitungsartikeln zu bewältigen, verwendet das *e-Identity*-Projekt diverse sprachtechnologische Werkzeuge. Das *e-Identity*-Projekt befindet sich aktuell in seinem letzten Projektjahr und hat bereits Tools und Verfahren entwickelt, die nun im Rahmen des Workshops der breiteren Forschungslandschaft der Digital Humanities im deutschsprachigen Raum präsentiert werden sollen. Neben der Präsentation unserer Forschungsergebnisse soll ein weiterer Schwerpunkt auf der Präsentation externer Forschungsprojekte liegen, die aktuell an der Schnittstelle von Computerlinguistik und Sozialwissenschaften durchgeführt werden.

Format der Workshops

Der zweitägige Workshop soll im Vorfeld zur Digital Humanities Jahrestagung in Graz an den Tagen 23. und 24. Februar 2015 stattfinden. Als Format für den Workshop schlagen wir zwei Phasen vor:

1) In einer ersten Phase erhalten die Workshop-Teilnehmer die Möglichkeit, über ihren Einsatz von Software-Werkzeugen oder anderen quantitativen (z.B. korpuslinguistischen) Methoden der Inhaltsanalyse großer Textmengen zu berichten. Sie sollen dabei konkret von ihrer Forschungspraxis im Rahmen ihrer sozialwissenschaftlichen Forschungsprojekte berichten. Ziel der Vorträge ist es, möglichst viele Einblicke in methodische und technische Einzelheiten der empirischen Analysen zu gewinnen, was auch die Demonstration von Software-Werkzeugen einschließt. Innerhalb dieses Blocks sollen auch die im *e-Identity*

Verbund entstandenen Werkzeuge präsentiert werden: Eine Explorations-Werkbank für die Konstruktion und manuelle Annotation von Korpora aus heterogenen Textquellen, und der *Complex Concept Builder* – eine mehrschichtige Analyse-Pipeline für die automatische Annotation der Texte mit Linguistik-naher Information. Insbesondere von Relevanz sind hier die Endprodukte der Pipeline, die darauf ausgelegt sind, von Sozialwissenschaftlern verwendet zu werden.

Das Format der ersten Workshop-Phase besteht aus jeweils 15 bis max. 20-minütigen Vorträgen (inklusive Demos) und anschließender 10-minütiger Diskussion. Insgesamt sollen ca. 5 –7 externe Forschungsgruppen die Möglichkeit erhalten, ihre Forschungsprojekte vorzustellen. Die Beitragenden werden von den Workshop-Organisatoren unmittelbar angesprochen – es wird also kein offizieller Call for Papers veröffentlicht. Nichtsdestotrotz wird der Workshop für weitere Forscherinnen und Forscher der e-Humanities geöffnet und über unterschiedlichste Kanäle, wie die Newsletter der FAG 8 der Clarin-D-Community publik gemacht.

2) Die Vorträge dienen dann in der zweiten Phase als Grundlage für eine breitere Reflexion und vergleichende Analyse der vorgestellten Werkzeuge und Resultate. Die TeilnehmerInnen werden aufgefordert, möglichst selbstkritisch und transparent über Schwierigkeiten und Herausforderungen ihrer methodischen Ansätze zu berichten. Dabei stehen u.a. der Vergleich der gesetzten Forschungsziele und die erreichte Funktionalität sowie eine Diskussion hinsichtlich der Übertragbarkeit auf unterschiedliche Anwendungsszenarien im Vordergrund. Ziel des zweiten Themenblocks besteht darüber hinaus in einer gemeinsamen Bestandsaufnahme von Erfahrungswerten der konkreten Zusammenarbeit von Sozialwissenschaften und Informatik-nahen Disziplinen:

- Was wurde bisher erreicht – was „funktioniert“ nunmehr?
- Was bedeutet es wenn etwas „funktioniert“? Während Informatik-nahe Disziplinen sich beispielsweise über eine 80%ige Trefferrate auf natürlichsprachlichem Text durchaus freuen, gibt es in den Sozialwissenschaften meistens ein anderes Verständnis von „funktionieren“. Wie kann hier eine Brücke geschlagen werden?
- Welche ursprünglichen Ziele oder Pläne haben sich als noch nicht umsetzbar erwiesen?
- Welche neuen Pläne oder Ziele ergeben sich aus Anstößen der bisherigen Zusammenarbeit?

Das Format des zweiten Themenblocks unterscheidet sich methodisch vom ersten Themenblock. Hier sollen keine Präsentationen stattfinden, sondern vielmehr moderierte Diskussionen, sowie kürzere Gruppenarbeits-Phasen. Die Beitragenden werden von den Workshop-Organisatoren unmittelbar angesprochen, es wird also keinen öffentlichen Call for Papers geben. Das bedeutet freilich nicht, dass es sich um einen "geschlossenen" Workshop handeln soll; im Gegenteil sind weitere Teilnehmer sehr willkommen. Die Vorträge werden so konzipiert, dass sie als Grundlage für die Diskussion in Phase 2 dienen können.

Je nach Lage der Interessen und Zusammensetzung der Teilnehmergruppe sind auch kurze Phasen der Gruppenarbeit denkbar.

Für den Erfolg dieses Szenarios erscheint es uns wichtig, dass beide Phasen nicht unmittelbar aufeinander folgen, sondern die Teilnehmer nach Abschluss der Präsentationen eine „Bedenkzeit“ haben, bevor die Diskussionsphase einsetzt. Im Unterschied zum eigentlich vorgegebenen System halbtägiger Pre-Conference Workshops schlagen wir daher vor, unseren Workshop an zwei halben Tagen stattfinden zu lassen: Phase 1 am Montagnachmittag, Phase 2 am Dienstagvormittag.

Zielgruppe des Workshops

Dieser Workshop richtet sich an andere Forschungsgruppen, die sich bereits im fortgeschrittenem Stadium ihres Projektes befinden. Sowohl Teilnehmer aus den Informatik-nahen Bereichen als auch aus den Sozialwissenschaften sind angesprochen.