Abstract: *Topic Modelling* des *Letters of 1916* Briefkorpus

Roman Bleier, Trinity College Dublin, Irland
Email: bleierr@tcd.ie

Einführung und Kontext

Das Jahr 1916 ist in der irischen Geschichte und Erinnerung eng mit dem Unabhängigkeitskampf gegen Grossbritannien verbunden und dem sogenannten Osteraufstandes, oder *Easter Rising*. Das hundertjährige Jubiläum des Osteraufstandes im Jahre 2016 wird national und international große Aufmerksamkeit erregen und mehrere Projekte zur Aufarbeitung der Geschichte des Jahres 1916 versprechen neue Erkenntnisse rund um den Osteraufstand. Das *Letter of 1916: Creating History* Projekt ist eines der umfangreichsten Unternehmen dieser Art.

Das *Letters of 1916: Creating History* Projekt hat als Ziel private Briefe, welche in den Monaten vor und nach dem Osteraufstand geschrieben wurden, zu sammeln und zu digitalisieren. Angestrebt wird ein Korpus von bisher unveröffentlichten, privaten Briefen aufzubauen, das die Studie der Situation in Dublin und Irland aus der Sicht der Bevölkerung ermöglicht. In den Worten der Projektleiterin Professor Schreibman:

> *'Through these letters we will to bring to life to the written word, the last words, the unspoken words and the forgotten words of ordinary people during this formative period in Irish history. All too often our emphasis is on the grand narrative focusing on key political figures. But as we approach the centenary of the Easter Rising we want to try to get a sense of how ordinary people coped with one of the most disruptive periods in contemporary Irish history...'*[1]

Ein weiteres Bestreben des Projekts ist es die Bevölkerung Irlands in das Sammeln und Transkribieren der Briefe einzubinden. Durch die Methode des *Crowdsourcing* wird nicht nur Arbeit ausgelagert, sondern auch Interesse am Projekt geweckt. Es ist das erste Crowdsourcing Projekt dieser Art in Irland und hat als Vorbild das Projekt *Transcribe Bentham* (UCL).

[1] Letters of 1916 Press Release, 27 September 2013, http://dh.tcd.ie/letters1916/wp-content/uploads/2013/09/1916-Letters-Press-Release-27-September-20131.pdf (09.11.2014).

Email: bleierr@tcd.ie

Die Webseite des *Letters of 1916* Projekts wurde im September 2013 eröffnet. Bilder von Briefe können dort hochgeladen, gelesen und transkribiert werden. Im Laufe des vergangenen Jahres sind auf diesem Wege über 1350 Briefe gesammelt worden, und in den nächsten zwei Jahren wird erwartet, dass das Korpus um das Doppelte oder Dreifache wachsen wird. Für 2016 ist geplant, dass das Korpus in Form eines Bild- und Textarchives der Öffentlichkeit zugängig gemacht wird. In Vorbereitung auf diese zweite Phase des Projekts experimentieren die Projektmitarbeiter mit Methoden zur Korpusanalyse und visuellen Veranschaulichung des Korpus. Eine dieser Methoden ist *Topic Modelling*.

Topic Modelling ist eine Technik die seit einigen Jahren in den Digitalen Geisteswissenschaften verwendet wird. Die Methode ist eng verknüpft mit dem Begriff des *Distant Reading* und *Macroanalysis*, dem Bestreben großer Textkorpora durch automatisierte Analyse und visuelle Aufbereitung mittels Diagrammen und Graphen Herr zu werden.[2] Vereinfacht ausgedrückt wird bei *Topic Modelling* versucht computergenerierte Themen aus einem Korpus von (meist) Textdokumenten zu gewinnen. Die einzelnen Dokumente des Korpus können je nach ihrer probabilistischen Nähe einem oder mehreren dieser Themen zugeordnet werden. In den Digitalen Geisteswissenschaften wird vor allem das LDA Model von Blei, Ng und Jordan verwendet, da es standardmäßig in Tools wie Mallet oder Gensim[3] implementiert ist.[4]

Das *Letters of 1916* Korpus wurde bisher zweimal durch *Topic Modelling* untersucht. Im Juni 2014 wurde eine erste Untersuchtung mit knapp 700 Briefen durchgeführt. Das Verfahren wurde im Januar 2015 mit 1350 Briefen wiederholt. Zur automatisierten Generierung von Themen wurde das *Topic Modelling Tool Mallet* verwendet. Sechzehn computergenerierte Themen wurden erstellt und die Briefe im Korpus je nach ihrer probabilistisch Nähe zu den einzelnen Themen in einem Graphen eingezeichnet. Für die graphische Aufbereitung wurde das Programm Gephi verwendet. Ziel dieser periodischen Untersuchung ist es, vor allem die Korpuszusammensetzung und Korpusentwicklung zu studieren und zu dokumentieren.

---

[2] Besonders Franco Moretti und Mat Jockers haben in den letzten Jahren zur Verbreitung dieser Methode in den Digital Humanities beigetragen. Franco Moretti, Distant Reading (2013). Mat Jockers, Macroanalysis: Digital Methods and Literary History (2013).

[3] Mallet: MAchine Learning for LanguagE Toolkit, http://mallet.cs.umass.edu/ (09.11.2014); Gensim: topic modelling for humans, http://radimrehurek.com/gensim/ (09.11.2014).

[4] Grundlegender Artikel zu LDA ist: Blei, et al., Latent Dirichlet Allocation, Journal of Machine Learning Research 3 (2003) 993-1022, http://machinelearning.wustl.edu/mlpapers/paper_files/BleiNJ03.pdf (09.11.2014).

Email: bleierr@tcd.ie

Beim Hochladen eines Briefes müssen bestimmte Metadaten angegeben werden. Name und Geschlecht des Absenders, das Datum des Briefes, und eine von sechzehn Kategorien, in die der Brief eingeordnet werden kann. Diese Kategorien sind von den Editoren zu Beginn des Projekts festgelegt worden und beinhalten Themen wie: Easter Rising, World War 1, Official Documents, Love Letters, Family Life, etc. Die von Menschenhand zugeordneten Themen sind besonders interessante Metadaten, da sie einen Vergleich mit den computer-generierten Themen ermöglichen.

Mein Vortrag wird einen Überblick über das Letters of 1916 Projekt und die Untersuchung des Korpus durch *Topic Modelling* bieten. Die Resultate der Untersuchen werden vorgestellt, und etwaige Probleme, die bei der Korpusreinigung und der Datenanalyse aufgetreten sind, werden diskutiert. Die Untersuchen des Teilkorpus soll feststellen, inwieweit eine Lesung des entgültigen Briefkorpus durch *Topic Modelling* möglich und sinnvoll ist und welche Erkenntnisse dadurch gewonnen werden können.