DGfS-CL

Fritz Kliche
Universität Hildesheim

e-Identity: Sprachtechnologische Werkzeuge und Methoden für die Analyse mehrsprachiger Zeitungstexte in den Sozialwissenschaften

Wir entwickeln Werkzeuge zur Aufbereitung rohen Textmaterials für die textanalytische Arbeit in den Digital Humanities. Unser Paket umfasst Werkzeuge zum Import, zur Ablage und Filterung der Texte in einem Repository und zur Exploration der Texte unter Verwendung von NLP-Techniken.

Ausgangspunkt sind Zeitungsdaten (mehrsprachiges Korpus, ca. 700.000 Artikel) aus unterschiedlichen Medienarchiven und mit unterschiedlichen Datenstrukturen. Wir entwickeln eine *Explorationswerkbank*, um die Daten in ein homogenisiertes Korpus zu überführen. Die Explorationswerkbank bereinigt die Daten um Textdubletten (doppelte Artikel) und Semi-Dubletten (ähnliche Artikel), findet Samplingfehler, interpretiert die den Texten beigefügten Metadaten und übernimmt die Aufgaben zur Konvertierung von Formaten (TXT, RTF, DOCX, HTML in XML) und Zeichenkodierungen (z. B. ISO-8859-1 in UTF-8).

Unsere Projektpartner aus der Sozialwissenschaft operieren mit abstrakten Begriffen, die nicht wörtlich in dieser Form in Zeitungsartikeln vorkommen: Sie suchen im *e-Identity*-Korpus Artikel, in denen kollektive (nationale, europäische, religiöse usw.) Identitäten zum Ausdruck kommen (vgl. Kantner 2011). Einfache Keyword-Anfragen genügen diesen Anforderungen nicht. Wir entwickeln stattdessen einen *Complex Concept Builder*, der über semi-automatische Verfahren (Active Learning) zur Klassifikation von Topics die Erstellung eines Samples anbietet. Er findet Unterstützung durch NLP-Methoden wie Named Entity Recognition, Sentiment-Analyse oder die Identifikation von direkter und indirekter Rede (vgl. Blessing et al. 2013).

Die Werkzeuge stellen Prototypen für Arbeiten zur Operationalisierung abstrakter Konzepte aus den Digital Humanities dar. Das Poster beschreibt Architektur und Komponenten der Explorationswerkbank und gibt Beispiele ihrer Anwendung.

Literatur

Blessing, A., Sonntag, J., Kliche, F., Heid, U., Kuhn, J., Stede, M., 2013. Towards a tool for interactive concept building for large scale analysis in the humanities, in: Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, Association for Computational Linguistics, Sofia, Bulgarien. pp. 55–64. URL: http://www.aclweb.org/anthology/W13-2708.

Kantner, C., 2011. European Identity as Commercium and Communio in Transnational Debate on Wars and Humanitarian Military Interventions, RECON Online Working Paper 2011/37, Arena Oslo, Oslo, Norwegen. URL: http://www.uni-stuttgart.de/soz/ib/mitarbeiter/kantner.html.