56. Jahrestagung des Leibniz-Instituts für Deutsche Sprache vom 10. bis 12. März 2020

DEUTSCH IN EUROPA

Sprachpolitisch, grammatisch, methodisch

METHODENMESSE

Mittwoch, 11. März 2020, 16:00 Uhr

WEB CRAWLING ZUR ANALYSE VON UNTERSCHIEDEN DER DEUTSCHEN SPRACHVARIETÄTEN IN EUROPA

Christopher Schröder, Thomas Eckart, Dirk Goldhahn, Uwe Quasthoff und Gerhard Heyer

Im Projekt Deutscher Wortschatz werden täglich zahlreiche Web- und Nachrichtenquellen mittels Web Crawling abgerufen und verarbeitet. Das Crawling dieser Quellen produziert drei Arten von Datensätzen: 1. In Zeitscheiben, um möglichst große und repräsentative Textmengen zu gewinnen. 2. Tagesaktuell, zur Untersuchung von relevanten Themen des jeweiligen Tages und deren Veränderung über die Zeit. 3. Zusätzlich werden die tagesaktuellen Korpora nach Jahresende zu einem großen Jahreskorpus kompiliert, welches unter anderem jeweils für die Domains .de, .at und .ch erstellt wird. Über eine Reihe von Vorverarbeitungsschritten werden die resultierenden Texte aufbereitet und mit statistischen Informationen angereichert. Als Ergebnis dieser Prozesse entstehen Korpora, die zur Analyse von Unterschieden in der Nutzung der deutschen Sprache in verschiedenen europäischen Ländern benutzt werden können. Die erzeugten Korpora werden sowohl auf tagesaktuelle Trends, als auch auf Unterschiede in Vorkommen und Häufigkeit von Wortformen untersucht.

Literatur:

D. Goldhahn, T. Eckart & U. Quasthoff: Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. In: Proceedings of the 8th International Language Ressources and Evaluation (LREC'12), 2012.