Datengetriebene Untersuchung latenter Muster distributioneller Semantik

Nur selten hat in der jüngeren Geschichte ein Ereignis Lebenswirklichkeiten so stark beeinflusst und das Wissen und die Erfahrungen der Menschen weltweit so geprägt und verändert wie die derzeit noch andauernde Corona-Pandemie. Es ist anzunehmen, dass sich mit dem sich ändernden Welt- und Standardwissen der Kommunizierenden auch deren Sprache verändert. Dies macht sich nicht nur in der Lexik bemerkbar, indem viele neue Begriffe in den aktiven Wortschaft der Sprecher wandern, sondern auch in der Semantik und der Pragmatik, wenn beispielsweise Begriffe aus einer Gebrauchsdomäne wie der Fachsprache der Medizin vermehrt in der Umgangssprache Verwendung finden.
Ziel des Projekts ist die datengetriebene Analyse dieser semantischen Veränderungen anhand von Konzepten und Begriffen in ihrem jeweiligen Kontext. Auf Grundlage der distributionellen Semantik sollen auf großen Korpora mit Hilfe aktueller Deep-learning-Methoden Modelle extrahiert werden, in denen die zu untersuchenden Muster latent vorhanden sind. Mit diesen Repräsentationen lassen sich lexikalische Einheiten und ihre semantischen Veränderungen in verschiedenen Kontexten, zu unterschiedlichen Zeitpunkten, in unterschiedlichen Sprachen oder Medien (sog. "alternative Medien", social Media) vergleichen.
In einer ersten Phase des Projekts werden Korpora (speziell aus dem Zeitraum der Corona-Pandemie) gesammelt, gesichtet und vorverarbeitet. Die Medienanalyseplattform AYLIEN bietet einen Datensatz von etwa 1,2 Mio. englischsprachigen Nachrichtentexten aus dem Zeitraum zwischen November 2019 und Mai 2020 (https://blog.aylien.com/free-coronavirus-news-dataset/). Eine erste Kookkurrenzanalyse mit Bigrammen zeigt anhand deren Häufigkeitsverteilungen bereits die starken sprachlichen Veränderungen in den englischsprachigen Medien über diesen Zeitraum der Pandemie: Ist beispielsweise zu Anfang des Jahres noch von einem „pneumonia outbreak“ die Rede, etabliert sich seit März der Begriff „coronavirus pandemic“, der auch andere Bezeichnungen wie „novel virus“ oder „new virus“ ablöst.
Als nächster Schritt werden Korpora erstellt und vorberarbeitet, die einen Vergleich zwischen der Zeit vor und der während der Pandemie und zwischen verschiedenen Sprachräumen ermöglichen sollen.