DONNERSTAG, 16. MÄRZ 2017, 11.30 UHR

QUANTITATIV-EMPIRISCHE ANSÄTZE ZUR ANALYSE LEXIKALISCHER DATEN. METHODENREFLEXION AM BEISPIEL VON WANDEL UND SEQUENZIALITÄT

Carolin Müller-Spitzer / Sascha Wolfer (IDS)

„The language looks different if you look at a lot of it at once."(Sinclair 1991: 100): Dies ist denken wir der Grundgedanke, auf dem die quantitativ ausgerichtete empirische linguistische Forschung aufbaut: Wir wollen große Mengen sprachlichen Materials auf einmal in den Blick nehmen und durch geeignete Analysemethoden sowohl neue Phänomene entdecken als auch bekannte Phänomene systematischer erforschen. Das Ziel unseres Vortrags ist es dabei weniger, einzelne solche Erkenntnisse vorzustellen, sondern anhand von zwei Themenbereichen methodisch zu reflektieren, wo der quantitativ empirische Ansatz wirklich so funktioniert wie erhofft, wo aber auch – vielleicht sogar systembedingte – Grenzen liegen und welche Fallstricke zu beachten sind.

Wir greifen zu diesem Ziel die zwei Themenbereiche Sprachwandel und Sequenzialität heraus. Am Beispiel der Analyse von Sprachwandel können wir illustrieren, wie bestechend auf der einen Seite der quantitativ empirische Ansatz zur Erforschung dieses Phänomens ist, wie schwierig seine Umsetzung allerdings in der Praxis sein kann. Drei Stichpunkte sind dabei: i) Die Zipf-Verteilung sprachlicher Daten, die dazu führt, dass fast alle lexikalischen Innovationsprozesse im Bereich der seltenen Sprachereignisse liegen und mit häufigkeitsbasierten Analysemethoden schwer von anderen Phänomenen sprachlicher Varianz zu trennen sind. ii) Die Datenlage: Viele Innovationsprozesse spielen sich nicht in der geschriebenen Sprache ab, wie sie in den großen verfügbaren Textkorpora erfasst sind. iii) Scheinkorrelationen: Je nachdem, welche statistischen Methoden man anwendet, sieht man Zusammenhänge, die sich bei genauerer Inspektion als nicht belastbar erweisen.

Am Beispiel der Sequenzialität lässt sich auf der anderen Seite zeigen, wie gewinnbringend es sein kann, mit quantitativen Methoden große Sprachmengen in den Blick zu nehmen und dabei die sequenzielle Organisation von Sprache, die in der Psycholinguistik mehr als in der Korpuslinguistik (Stichwort Bag-of-words-Ansatz) immer eine tragende Rolle gespielt hat, auf die Analyse von geschriebenen Daten zu übertragen. Wir zeigen hier, dass wir anhand eines großen Bibelkorpus mit über 1.500 Bibelübersetzungen in über 1.200 Sprachen nachweisen konnten, dass es in allen untersuchten Sprachen einen trade-off zu beobachten gibt zwischen den Informationen, die über die Wortstellungs- vs. über die Wortstrukturregularität vermittelt werden: Jene Sprachen, die viel Information über die Wortstellung vermitteln, übertragen umso weniger über die Wortstruktur und andersherum.

Wir hoffen insgesamt mit unserem Vortrag einen Beitrag dazu zu leisten, auf der einen Seite die vielfältigen Möglichkeiten der quantitativ-empirischen Herangehensweise zu sehen, aber auf der anderen Seite auch die Grenzen dieses Ansatzes, insbesondere was die Datenlage angeht, zu reflektieren und damit über die Interpretationskraft der Verfahren diskutieren zu können.