58. Jahrestagung des Leibniz-Instituts für Deutsche Sprache vom 15. bis 17. März 2022

KORPORA IN DER GERMANISTISCHEN SPRACHWISSENSCHAFT – MÜNDLICH, SCHRIFTLICH, MULTIMEDIAL

Donnerstag, 17. März 2022, 11:05 Uhr

Korpusaufbau zwischen Standard und Innovation (am Beispiel von GiesKaNe)

Volker Emmrich / Mathilde Hennig (Universität Gießen)

Die Auseinandersetzung mit Standards ist zentral für jedes Korpusprojekt und muss unter Aspekten wie Forschungsinteresse, wissenschaftliche Originalität, eristischer Diskurs und Usability betrachtet werden. In einer solchen kritischen Auseinandersetzung wird deutlich, dass Standards einerseits im Konflikt mit Grundprinzipien wissenschaftlicher Forschung wie Innovation und Originalität stehen, während sie andererseits unter Gesichtspunkten wie Usability und Vergleichbarkeit von Daten unverzichtbar für die Einbettung eines neuen Korpus in die Korpuslandschaft sind. Beide Aspekte müssen aber keineswegs als einander ausschließende Alternativen betrachtet werden, da über die Möglichkeit der Mehrebenenannotationen die datenstrukturellen Voraussetzungen für einander ergänzende innovative und standardkonforme Annotationen gegeben sind und durch die Methoden des maschinellen Lernens eine effiziente Ergänzung möglich sein sollte: Wenn qualitative hochwertige komplexe Annotationen nach einem Annotationsschema vorliegen, können diese als unabhängige Variablen bzw. als Eingabe für die Vorhersage der Analyse nach einem anderen Annotationsschema dienen.

Überlegungen dieser Art werden im Vortrag am Beispiel von GiesKaNe (= GießenKassel-Neuhochdeutsch) präsentiert. Das aktuell im Rahmen eines DFG-Langfristvorhabens an den Standorten Kassel (Leitung: Vilmos Ágel) und Gießen (Leitung: Mathilde Hennig) im Aufbau befindliche Korpus arbeitet im Rahmen der Zielsetzung der Erschließung syntaktischer Grundstrukturen des Neuhochdeutschen mit innovativen Ansätzen, die der Komplexität der Annotationsziele (syntaktische Grundstrukturen von Text über Satz bis hin zu Wortgruppe und Wort, systematische Form-Funktionszusammenhänge, Syntax-Semantik-Beziehungen) gerecht werden sollen. Herzstück ist ein eigener Baumbankansatz, aber auch das Wortarttagging folgt einem neuen Modell. Das Ziel des geplanten Vortrags besteht darin, am Beispiel der Wortartannotation (GiesKaNe vs. STTS und HiTS) zu zeigen, dass mit Innovationen dieser Art einerseits neue Akzente in der Korpuslinguistik gesetzt werden können, dass sie aber nicht in eine idiosynkratische Einbahnstraße führen müssen, wenn man die im vorliegenden Modell vorhandenen Annotationen systematisch für die Rekonstruktion anderer Modelle nutzt und diese in der Mehrebenenannotation ergänzend für die Korpusnutzung anbietet.