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

DEUTSCH IN EUROPA

Sprachpolitisch, grammatisch, methodisch

Mittwoch, 11. März 2020, 9:00 Uhr

MULTILINGUALE SPRACHRESSOURCEN FÜR DIE LINGUISTISCHE FORSCHUNG

Erhard Hinrichs (IDS / Universität Tübingen)

Sprachressourcen in digitaler Form liegen für ein immer breiteres Spektrum von Einzelsprachen vor. Linguistisch annotierte Korpora ermöglichen es, gezielt nach linguistischen Mustern auf der Wort-, Phrasen-, und Satzebene zu suchen und in quantitativer und qualitativer Hinsicht auszuwerten. Mit dem Brown Corpus und dem Lancaster-Oslo-Bergen (LOB) Corpus lagen bereits in den 1960er und 1970er Jahren entsprechende Ressourcen für das amerikanische bzw. das britische Englische vor. Inzwischen hat sich nicht nur der Umfang derartiger Ressourcen signifikant vergrößert, sondern auch die Anzahl von Sprachen, für die sie zur Verfügung stehen. Wurden die frühen Korpora wie das Brown und LOB Corpus durch rein manuelle Annotationen angereichert, so lassen sich große Textmengen inzwischen mit Hilfe von computerlinguistischen Verfahren (teil-)automatisiert annotieren, allerdings mit entsprechenden Einbußen bei der Qualität der Annotationen.

Ab Mitte der 1980er Jahre wurden Textkorpora über die Annotation von Einzelwörtern hinaus auch durch syntaktische Informationen angereichert. Diese sog. Baumbanken wurden zunächst zum Trainieren von statistischen Parsern verwendet. Sie stellen aber bei hinreichender Größe auch eine interessante Ressource für linguistische Analysen dar. Die den Baumbanken zugrundeliegenden syntaktischen Analysen beruhen auf verschiedenen Syntaxtheorien der generativen Grammatik, der Phrasenstrukturgrammatik und der Dependenzgrammatik. Um sprachvergleichende Untersuchungen in Baumbanken unterschiedlicher Einzelsprachen zu erleichtern, hat sich das Kooperationsprojekt *Universal Dependencies* konstituiert mit dem Ziel, Baumbanken mit vergleichbaren Dependenzannotationen für möglichst viele Sprachen der Welt zu erstellen.