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

KORPORA IN DER GERMANISTISCHEN SPRACHWISSENSCHAFT – MÜNDLICH, SCHRIFTLICH, MULTIMEDIAL

Dienstag, 15. März 2022, 13:45 Uhr

Das Gesamtkonzept des Deutschen Referenzkorpus DeReKo – Vom Design bis zur Verwendung und darüber hinaus

Marc Kupietz, Harald Lüngen & Nils Diewald (IDS)

Aufgabe und Ziel des Deutschen Referenzkorpus DeReKo ist es, eine allgemeine Forschungsdatengrundlage für das IDS selbst und für die synchron arbeitende germanistische Linguistik insgesamt dauerhaft zu sichern. Zu diesem Zweck wird DeReKo seit 1968 am IDS aufgebaut und laufend stichprobenartig um ein möglichst breit gefächertes Spektrum des aktuellen deutschen Schriftsprachgebrauchs erweitert. LinguistInnen und, sofern möglich, auch Forschende aus angrenzenden Disziplinen sollen durch DeReKo in die Lage versetzt werden, sich für eine große Bandbreite an Fragestellungen und Sprachdomänen geeignet stratifizierte Sub-Korpora zu definieren, mithilfe derer sie bestehende Hypothesen zuverlässig testen und interessante neue Hypothesen gewinnen können.

Wir wollen in unserem Beitrag die sich durch die einschlägige Thematik der IDS-Jahrestagung bietende Gelegenheit nutzen, um einen Blick hinter DeReKo's Kulissen zu gewähren. Zum einen werden wir dabei die strategische Konzeption des Korpusausbauprojekts im Kontext seiner angrenzenden Projekte beleuchten. Dabei werden wir nach einer kurzen Rekapitulation von DeReKo's Grundlagen und Designprinzipien näher auf Erweiterungsstrategien eingehen und unseren Ansatz vorstellen, DeReKo mit begrenzten Ressourcen für einerseits möglichst viele, andererseits aber auch für innovative und anspruchsvolle Anwendungen nutzbar zu machen. Am Beispiel der Tokenisierung und der (Re-)Konstruktion und Modellierung von Metadaten werden wir außerdem exemplarisch Herausforderungen diskutieren, die häufig bei der Kompromissfindung zwischen unterschiedlichen linguistischen Anforderungen und allgemeinen wissenschaftlichen Maximen vor dem Hintergrund ökonomischer und technischer Notwendigkeiten entstehen. Dies erscheint uns relevant, da solche Zielkonflikte und notwendigen Kompromisslösungen oft schwer dokumentierbar sind und bei der Korpusnutzung unserer Erfahrung nach selten erahnt werden und daher einen oft nicht bemerkten aber potenziell starken Einfluss auf Untersuchungsergebnisse haben können. Zudem zeigt sich dabei, wie sehr auch bei solchen vermeintlichen Detailentscheidungen, die im Zuge der Erstellung großer Allzweck-Korpora häufig gefällt werden müssen, heterogene sprachwissenschaftliche, technische und infrastrukturelle Aspekte miteinander in Wechselwirkung stehen und welche Konsequenzen bzgl. der Verwendbarkeit damit verbunden sind.

Neben diesem Blick hinter die DeReKo-Kulissen wollen wir auch über aktuelle Neuerungen bei DeReKo und seinem Umfeld berichten. Diese betreffen zum einen substanzielle Erweiterungen von DeReKo selbst, wie z.B. im Bereich IBK und Social-Media das in Nottingham entwickelte NottDeuYTSch-Korpus mit YouTube-Kommentaren und ein am IDS entwickeltes und kontinuierlich erweitertes Twitter-Korpus sowie im Bereich der Fachsprache das in Greifswald und Leipzig aufwändig entwickelte Geschriebene Ingenieurwissenschaftliche Korpus Gingko. Zum anderen betreffen diese auch neue Verwendungsmöglichkeiten von De-

ReKo hinsichtlich vergleichender und kontrastiver Forschung im Kontext der Initiative Europäisches Referenzkorpus (EuReCo) und der Durchführung komplexer, mehrgliedriger, reproduzierbarer und replizierbarer Analysen mit Hilfe der Korpusanalyseplattform KorAP sowie seiner Bibliotheken für die Programmiersprachen R und Python.