DeReKo durchbricht Drei-Milliarden-Grenze

von Marc Kupietz

Mit dem schriftsprachlichen Korpusarchiv Deutsches Referenzkorpus (DeReKo) stellt das IDS eine empirische Datengrundlage für die linguistische Forschung im Umfang von jetzt mehr als 3,2 Milliarden Textwörtern zur Verfügung (umgerechnet ca. 8 Millionen Buchseiten). Damit umfasst DeReKo für jedes einzelne Jahr seit 1996 allein mehr Textmaterial als das wohl bekannteste englischsprachige Korpus (BNC) insgesamt.

Gegenüber dem Vorjahr konnte DeReKo, das u.a. über COSMAS II recherchierbar ist, um etwa 35% erweitert werden. Aus urheber- und lizenzrechtlichen Gründen ist leider nur ein Teil der archivierten Korpora außerhalb des IDS zugänglich. Jedoch wurde dieser öffentliche Teil gegenüber dem Vorjahr auf jetzt 2,2 Milliarden Wörter nahezu verdoppelt. Weitere Informationen zum aktuellen Bestand des Korpusarchivs finden Sie unter: <http://www.ids-mannheim.de/kl/projekte/korpora/archiv.html>.

Weiteres Wachstum und breitere Streuung avisiert

Das DeReKo-Projekt bemüht sich um einen stetig zunehmenden Umfang und eine immer breitere Streuung der Korpusdaten. Diese Zielrichtung entspricht dem Auftrag des IDS, den Gebrauch der deutschen Sprache kontinuierlich zu dokumentieren und wird zudem den Anforderungen der empirisch-linguistischen Forschung gerecht. Zum einen führen Korpusuntersuchungen auf der Basis eines größeren und breiter gestreuten Korpusarchivs offenkundig zu aussagekräftigeren Ergebnissen. Zum anderen ermöglicht DeReKo erst durch seine Größe die Anwendung statistischer korpuslinguistischer Methoden, die systemische Zusammenhänge im Sprachgebrauch aufdecken (z.B. Kookkurrenzanalysen oder multidimensionale Korpusanalysen).

Für einen empirisch fundierten Erkenntnisgewinn ist es zudem dringend ratsam, als Datengrundlage ein Korpus zu verwenden, das in Hinblick auf die jeweilige Grundgesamtheit als hinreichend repräsentativ oder ausgewogen angenommen werden kann. Diese Grundgesamtheit hängt jedoch erheblich von der linguistischen Fragestellung ab. Daher liegt es in der Verantwortung der empirischen Linguisten, sich aus dem Korpusarchiv nach bestimmten Kriterien ein für ihre jeweilige Fragestellung maßgeschneidertes Korpus (ein sogenanntes virtuelles Korpus) zusammenstellen zu lassen. Erst mit einem möglichst umfangreichen und breit gestreuten Korpusarchiv wird diese Vorgehensweise für ein breiteres Spektrum linguistischer Fragestellungen praktikabel.