DeReKo-Archiv jetzt mit fünf Milliarden Textwörtern

Zum größten digitalen Textarchiv für deutsche Texte der Gegenwart

von Harald Lüngen

Seit September 2011 umfasst das digitale Korpusarchiv Deutsches Referenzkorpus (DeReKo) des Instituts für Deutsche Sprache über 5 Milliarden Textwörter (bei Zählung jeder Wortform im laufenden Text). Das entspricht über 12,5 Millionen Buchseiten, wenn man durchschnittlich 400 Wörter pro Seite zugrunde legt. DeReKo enthält Zeitungstexte, Belletristik, Fachtexte und weitere Textsorten und wird seit dem Jahr 1964 fortlaufend ausgebaut. Das Archiv trägt dem Auftrag des IDS Rechnung, den Gebrauch der deutschen Gegenwartssprache fortlaufend zu dokumentieren. Es wird von Germanisten und Sprachwissenschaftlern in aller Welt genutzt, beispielsweise beim Verfassen von Grammatiken und Wörterbüchern. Eine DeReKo-Recherche kann dabei helfen, zu entscheiden, ob eine Wortverbindung oder ein Neuwort wie *Raucherkneipe* überhaupt häufig genug vorkommt, um seine Aufnahme in ein allgemeinsprachliches Wörterbuch zu rechtfertigen, ob das weibliche oder das neutrale Geschlecht für *Mail*, also *die Mail* oder *das Mail*, verbreiteter ist oder ob es regionale Unterschiede der Verwendung z. B. zwischen Deutschland, Österreich oder der Schweiz gibt. DeReKo wird auch eingesetzt für die Entwicklung neuartiger empirisch-sprachwissenschaftlicher Analysemethoden, welche auf eine möglichst große, um nicht zu sagen riesige Datengrundlage angewiesen sind.

Das Deutsche Referenzkorpus ist damit das größte digitale Textarchiv für deutsche Texte der Gegenwart. Es ist z. B. 50 mal größer als das bekannte englischsprachige „British National Corpus“, welches allerdings seit 1994 abgeschlossen ist und nicht mehr erweitert wird. Gegenüber 2010 konnte DeReKo noch einmal um ca. 20% ausgebaut werden. Einen großen Anteil an der jetzigen Erweiterung trägt die Aufbereitung und Integration aller deutschsprachigen Artikel und Autor-Diskussionen der Online-Enzyklopädie Wikipedia (811 Millionen Textwörter) bei. Neu ist aber auch ein erster Anteil von neuen belletristischen Werken und Fachbüchern, für die 2011 Nutzungsrechte von Buchverlagen eingeworben wurden (wobei das Gros dieser neu akquirierten Texte sich erst in den nachfolgenden DeReKo-Ausgaben niederschlagen wird).

DeReKo ist einem Urstichproben-Design verpflichtet. Das heißt, dass DeReKo als Korpusarchiv nicht als ausgewogene oder gar repräsentative Stichprobe der deutschen Gegenwartssprache konzipiert wurde, da es zweifelhaft ist, ob eine solche überhaupt sinnvoll definiert werden kann. Vielmehr stellen sich Forscher für ihre Fragestellungen aus der sehr großen Urstichprobe spezifische Teil-Stichproben in sogenannten virtuellen Korpora zusammen. Ein solches virtuelles Korpus kann anhand sprachlicher oder außersprachlicher Merkmale so definiert werden, dass es ausgewogen und repräsentativ in Bezug auf die Fragestellung ist. Beispielsweise kann für die Ermittlung und Untersuchung von neuen Wörtern und Wortverbindungen in der Sprache ein virtuelles Korpus definiert werden, das für jedes Jahr der letzten Dekade, also 2001, 2002, 2003, ..., 2010 etwa gleich viel Wortmaterial, z. B. jeweils zehn Millionen Wörter enthält.

Aus urheber- und lizenzrechtlichen Gründen ist ein Teil von DeReKo weiterhin nur am IDS nutzbar. Der überwiegende Teil (ca. 80%, d.h. knapp vier Milliarden Textwörter) ist aber zu wissenschaftlichen Zwecken weltweit über die Web-Schnittstelle Cosmas-II recherchierbar.

Wir danken herzlich allen Text- bzw. Lizenzspendern, ohne die DeReKo nicht möglich wäre! Weitere Informationen zum Bestand von DeReKo finden Sie unter: <www.ids-mannheim.de/kl/projekte/korpora/archiv.html>.

Der Autor ist wissenschaftlicher Mitarbeiter am Institut für Deutsche Sprache in Mannheim.