COSMAS II – Recherchieren in den Korpora des IDS

von Franck Bodmer

Mit knapp zwei Milliarden Wörtern bilden die Korpora des IDS die weltweit größte Sammlung elektronischer Korpora mit geschriebenen deutschsprachigen Texten aus der Gegenwart und der neueren Vergangenheit. Sie stellen für die Forschung eine unverzichtbare empirische Basis dar - nicht nur für die linguistische Forschung am IDS, sondern auch für die nationale und internationale Germanistikforschung, wie die ständig zunehmende Zahl an Online-Nutzern belegt. Sie werden aber auch in wachsendem Maße für interdisziplinäre Untersuchungen genutzt, etwa in den Fachgebieten Psychologie, Neurologie, Kognitionswissenschaft, Sprachtherapie, Kommunikations- und Medienwissenschaft und Statistik.[1]
Nur ein sehr begrenzter Anteil der IDS-Textsammlung kann für eigene Untersuchungen als Datei erworben werden. Schon wegen ihres Umfangs lassen sich die IDS-Korpora ohnehin nicht mehr ohne die Unterstützung einer spezialisierten Software verwalten, recherchieren und analysieren. Zu diesem Zweck wurden am IDS die Grundzüge einer Software namens *COSMAS* (*Corpus Search, Management and Analysis System*) zur Unterstützung korpusbasierter Forschung konzipiert und in Form von *COSMAS I* (Einsatz: 1992 bis März 2003) und *COSMAS II* (Einsatz: seit 2002) umgesetzt und eingesetzt.

Die IDS-Korpora werden in *COSMAS II* in vier bis fünf Archiven[1] als Einzelkorpora oder virtuelle Korpora verfügbar gemacht. Wer für seine Untersuchung einen speziellen Querschnitt durch die Texte braucht, kann sich in *COSMAS II* ein oder mehrere passende virtuelle Korpora selbst definieren und für spätere *COSMAS II*-Sitzungen sichern.

Die Suchanfragesprache ist für die Formulierung von linguistisch motivierten Recherchen konzipiert. Mit Hilfe des *Wort-*, *Satzabstands-* und *Grundform-operators* lassen sich genaue Muster und eine hohe Ausbeute erzielen.

Die Ergebnisse und Treffer werden dank der Informationen, die in den kontinuierlich gepflegten Korpora vorhanden sind, gut dokumentiert.

Mit der *Kookkurrenzanalyse*[2], die ebenfalls am IDS entwickelt wurde, wird dem Benutzer ein Werkzeug zur Hand gegeben, mit dem er die oft sehr langen

KWICs (KWIC = Keyword in Context) nach sprachlichen Mustern analysieren lassen kann.

COSMAS II ist ein kostenloser Internet-Dienst des IDS: Er steht sowohl institutsinternen als auch externen Nutzern in zwei Versionen zur Verfügung: als WINDOWS- und als Web-Version. Voraussetzung für die Benutzung ist die Registrierung und die Einhaltung der damit verbundenen Nutzungsrechte.[3]

Im Folgenden werden die beiden Benutzeroberflächen von *COSMAS II* präsentiert.

Die WINDOWS-Benutzeroberfläche

Die WINDOWS-Benutzeroberfläche ist ein Programm, das sich leicht von der Installationsseite[4] herunterladen und installieren lässt. Installationsanweisungen und Einstellung einer lokalen Firewall auf *COSMAS II* finden Sie ebenfalls auf dieser Seite. Im Falle von Problemen können Sie die *häufig gestellten Fragen*[5] konsultieren oder uns anschreiben.[6]

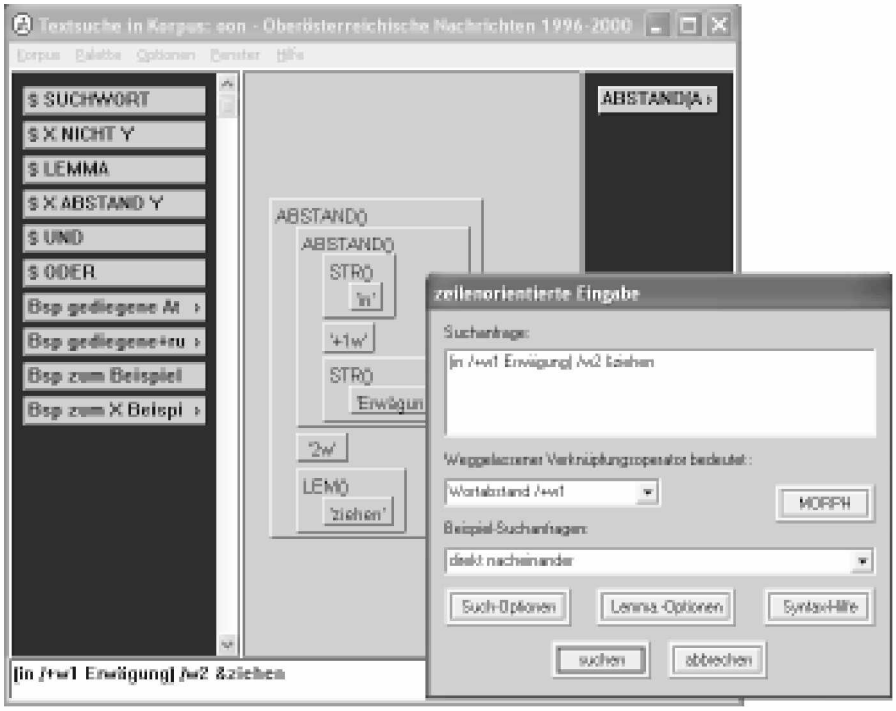

Abbildung 1: Das Textsuchfenster in der WINDOWS-Oberfläche

Es wird empfohlen, als erstes die kurze Übersicht über die Arbeitsabläufe in *COSMAS II* durchzulesen.[7] Nach dem Start der Software und der Anmeldung sind zudem Tipps und Erklärungen (die gelben Fenster) des *COSMAS II-Begleiters* eingestreut, die Sie im Umgang mit der Software unterstützen sollen (und die Sie jederzeit deaktivieren können).

Im Textsuchfenster (siehe Abbildung 1) entscheiden Sie, ob Sie die Suchanfragen mit Hilfe der Zeileneingabe (Menü *Fenster/Zeileneingabe* oder Tastenkombination *<alt>E*) oder des grafischen Assistenten (oder beider) formulieren möchten. Eine im Zeileneingabefenster editierte Suchanfrage wird automatisch auch in eine äquivalente grafische Suchanfrage umgewandelt, die Sie abändern oder ergänzen können. Sehr hilfreich ist der Einsatz des *Lemmatisierungsoperators,* der für eine Grundform u.a. die im ausgewählten Korpus vorkommenden Flexionsformen und Komposita erzeugt. Falls Sie sich im Archiv *TAGGED* befinden, in welchem die Texte morpho-syntaktisch annotiert sind, können Sie mit Hilfe des Assistenten *MORPH* Wortklassenannotationen in Ihre Suchanfrage einfügen.

Die Ergebnisse werden in unterschiedlichen Formen nacheinander präsentiert: als Tabelle von Statistiken (Anzahl Treffer pro Dokument / Korpus / Jahrgang / Jahrzehnt etc.); als KWIC (unsortiert, alphabetisch oder chronologisch sortiert); als Volltext.

Ein Schwerpunkt der Auswertung des KWICs ist die auf einem statistischen Verfahren beruhende *Kookkurrenzanalyse* (*KA*), die Sie im Ergebnisfenster aufrufen können und die auf-fallende Wortverbindungen in Form von *Kookkurrenz-Clustern* und *syntagmatischen Mustern* liefert. *COSMAS II* präsentiert die Ergebnisse der *KA* auf Wunsch in *synoptischer* Darstellung (siehe Abbildung 2), bei der sich das KWIC-Fenster an das Fenster mit den *KA*-Ergebnissen andockt.

COSMAS II-Web

Die Benutzeroberfläche *COSMAS II-Web* ist in erster Linie für die Korpusnutzer gedacht, die über kein WINDOWS-Betriebssystem verfügen. Dieser Zugang läuft in Ihrem WWW-Browser[8] und wird unter der URL http://www.ids-mann-heim.de/cosmas2-web/ ge-funden. Als Webanwendung hat *COSMAS II-Web* einen konzeptionellen Vorteil gegenüber der WINDOWS-Version: Sie brauchen nichts zu installieren, Sie müssen auch keine Updates einspielen.

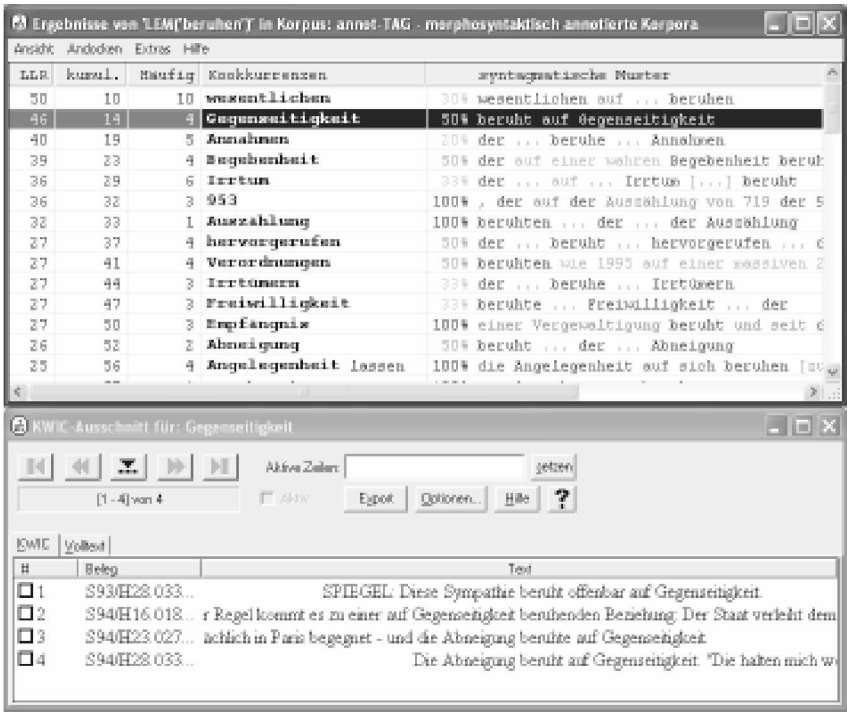

Abbildung 2: Synoptische Darstellung der Kookkurrenzanalyse mit KWIC-Fenster

COSMAS II-Web läuft immer mit der aktuellsten Version.

Die erste Version von *COSMAS II-Web* wurde Ende Mai 2005 freigegeben. Sie umfasst ein Minimalangebot an Funktionalitäten von der Archiv- und Korpusauswahl bis zum Export der Ergebnisse. Im Laufe der nächsten Monate wird der Funktionsumfang der Webversion sukzessive an den der WINDOWS-Version angeglichen.[9]

Welche Version Sie einsetzen möchten, hängt nun entweder von Ihrem Betriebsystem oder Ihrer Präferenz ab. Prinzipiell unterscheiden sich die beiden Programme durch die Benutzerführung und die Handhabung, nicht durch inhaltliche Konzepte: Sie greifen bei beiden auf dieselben Archive sowie auf dieselben vor- und selbstdefinierten Korpora zu; die Suchanfrage-syntax (in der Web-Version zurzeit nur die zeilenorientierte) ist in beiden Programmen dieselbe; gleiche Suchanfragen ergeben bei gleichen Einstellungen (bis auf die Zufallsauswahl) identische Treffermengen; Ihre Benutzerkennung und -einstellungen werden von beiden Programmen gemeinsam verwaltet.

Nach der Vervollständigung der Web-Version sollen aber weiterhin beide, die WINDOWS- und die Web-Version, parallel betrieben werden. Die Vorteile der WINDOWS-Version liegen darin, dass die Gestaltungsmöglichkeiten und die Benutzerführung verlässlicher und ausgereifter programmiert werden können.

COSMAS II-Web verdeutlicht durch sein Layout, welche Schritte aufeinander folgen. Man schreitet von Seite zu Seite voran, indem man auf die Schalter oder bei Tabellen auf einen Eintrag klickt. Die linke Menü- bzw. Navigationsleiste gibt wieder, welche Seiten bereits durchschritten wurden und welche bis zum Export noch bevorstehen. Diese Leiste dient sowohl dem Zurückspringen zu einer bestimmten Seite (z.B. zur Suchanfrage-Seite, um eine neue Anfrage zu formulieren, oder zur Korpus-Seite, um das Korpus zu wechseln) als auch dem Vorwärtsspringen. Der Status-bereich gibt auf jeder Seite Auskunft über die Wahl des Archivs und des aktiven Korpus, die aktive Suchanfrage und die Anzahl der Treffer.

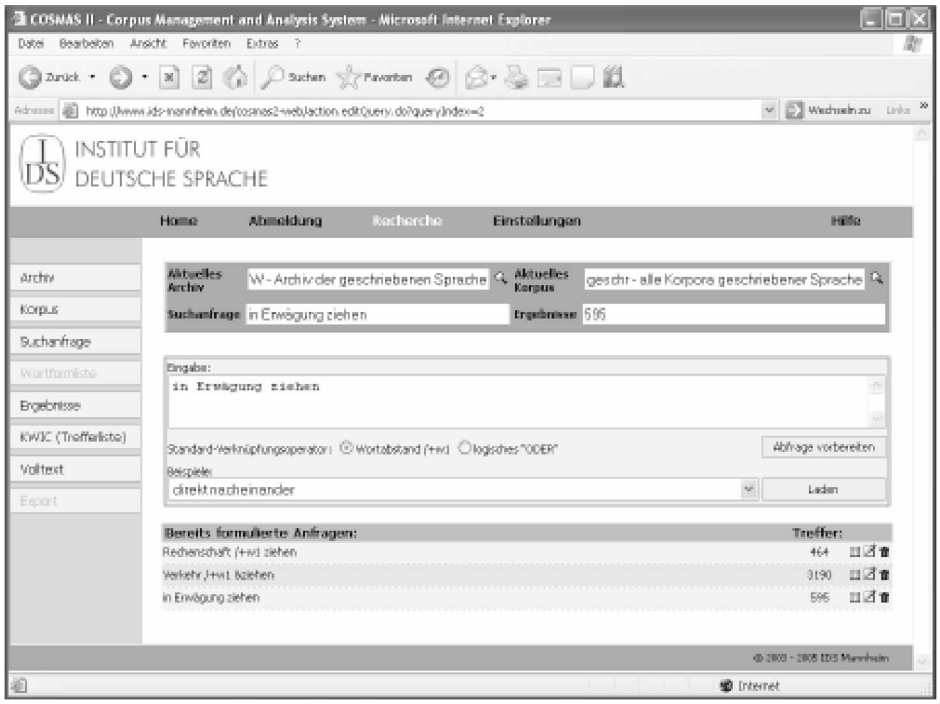

Abbildung 3: Die Suchseite in *COSMAS II-Web*

Aktuelle Korpusarbeiten, die sich in *COSMAS II* niederschlagen

Eine beständige Hintergrundaktivität ist die Korpusakquisition, -verbesserung und -anreicherung, die vom Projekt *Ausbau und Pflege der Korpora geschriebener Sprache* durchgeführt wird.[10] Ein- bis

zweimal im Jahr werden die Änderungen in die von *COSMAS II* angebotenen Archive übernommen. Dazu gehört insbesondere die laufende Ergänzung der Archive bzw. Korpora mit Texten aus dem letzten und dem laufenden Jahr. Dazu kommen in diesem Jahr z.B. die Eliminierung von überflüssigen oder doppelt vorhandenen Texten.

Zukünftiges

Neben der Ergonomie der beiden Benutzeroberflächen, die wir stets zu verbessern bemüht sind (über Ihre Rückmeldungen an *cosmas2@ids-mannheim.de* freuen wir uns), werden die bestehenden Funktionalitäten ergänzt und neue integriert. Außerdem fließen auch neue Daten in Folge der Aufbereitung und Anreicherung der IDS-Korpora in *COSMAS II* ein. Hier seien einige Punkte erwähnt:

Die Aufbereitung der IDS-Korpora erfährt zurzeit durch die Konvertierung in das *CES*[11]-Format eine Verfeinerung der Kodierung der Textstruktur und der auffindbaren Annotationen. Dadurch wird es in Zukunft möglich sein, u.a. Recherchen gezielt auf bestimmte Textbereiche einzuschränken (z.B. um Treffer in Überschriften auszuschließen). Das Archiv *TAGGED*, in welchem ein kleiner Ausschnitt aus dem Hauptarchiv morpho-syntaktisch annotiert wurde, soll auf weite Sicht ergänzt und hinsichtlich der Qualität der Annotationen verbessert werden.

Die virtuellen Korpora sollen mit Korpusstatistiken dokumentiert werden. Die Möglichkeiten, eigene virtuelle Korpora zusammenzustellen, sollen durch Hinzunahme weiterer bibliografischer Angaben wie z.B. Entstehungsdatum oder Textklasse erweitert und verfeinert werden.

Die Wortformlisten, die mit dem Lemmatisierungs- und Platzhalteroperator generiert werden, sollen mit Filtern ausgestattet werden, um kürzere und präzisere Listen zu erzeugen.

Die Angaben der absoluten Häufigkeiten der Suchergebnisse sollen um Angaben von relativen Häufigkeiten ergänzt werden. Auf das KWIC sollen Suchfunktionen und Filter eingesetzt werden, um spezifische Muster schneller zu lokalisieren.

Die Präsentation der Belege (im Volltext) soll durch zusätzliche Informationen aus den Quellen ergänzt werden.

Anmerkungen

[1] Mehr über die IDS-Korpora finden Sie unter: http://www.ids-mannheim.de/kt/projekte/korpora/

[2] Siehe die Korpus-Dokumentation unter http://www.ids-mannheim.de/cosmas2/referenz/

[3] Mehr über die Kookkurrenzanalyse und die damit erzeugte Kookkurrenzdatenbank erfahren Sie unter http://www.ids-mannheim.de/kt/projekte/methoden/ka.html

[4] Registrierung unter http://www.ids-mannheim.de/cosmas2/registrierung/

[5] Installationsseite von *COSMAS II*: http://www.ids-mannheim.de/cosmas2/install/

[6] FAQ von *COSMAS II*: http://www.ids-mannheim.de/cosmas2/hilfe/allgemein/faq.html

[7] Fragen und Anregungen richten Sie bitte an cosmas2@ids-mannheim.de

[8] Arbeitsabläufe in *COSMAS II*: http://www.ids-mannheim.de/cosmas2/hilfe/allgemein/quicktour.html

[9] Welche Browser auf welchen Plattformen unterstützt werden, finden Sie auf der Startseite von *COSMAS II-Web*.

[10] Welche Funktionalitäten in der Web-Version fehlen bzw. neu dazugekommen sind, finden Sie ebenfalls auf der Startseite von *COSMAS II-Web*.

[11] Siehe: http://www.ids-mannheim.de/kt/projekte/korpora/

[12] CES: Corpus Encoding Standard; siehe http://www.cs.vassar.edu/CES/