Franck Bodmer Mory

MIT COSMAS II »IN DEN WEITEN DER IDS-KORPORA UNTERWEGS«

Einleitung

COSMAS II (= Corpus Search, Management and Analysis System) ist nach COSMAS I und REFER die 3. Generation von Korpus-Recherchesystemen, die am IDS entwickelt und eingesetzt wird, um die linguistische Erforschung der eigenen umfangreichen Korpora im Haus und aus der ganzen Welt zu ermöglichen.

Immer wieder erhalten wir die Anfrage, ob die IDS-Korpora heruntergeladen werden können, um darin auf dem eigenen Rechner recherchieren zu können. Abgesehen davon, dass wir dadurch gegen die mit den Textgebern ausgehandelten Nutzungsbedingungen verstoßen würden, kämen auf den Nutzer unzählige Schwierigkeiten zu. Die DEREKO[1]-Korpora sind in dekomprimiertem Format einige Tera-Bytes (einige Tausende Giga-Bytes) groß; das Dekomprimieren allein beansprucht auf einem leistungsstarken Server Stunden, das Indexieren der Korpora in ein schnelles Datenbankformat einige Tage und das Entwickeln einer mit korpuslinguistischen Funktionalitäten

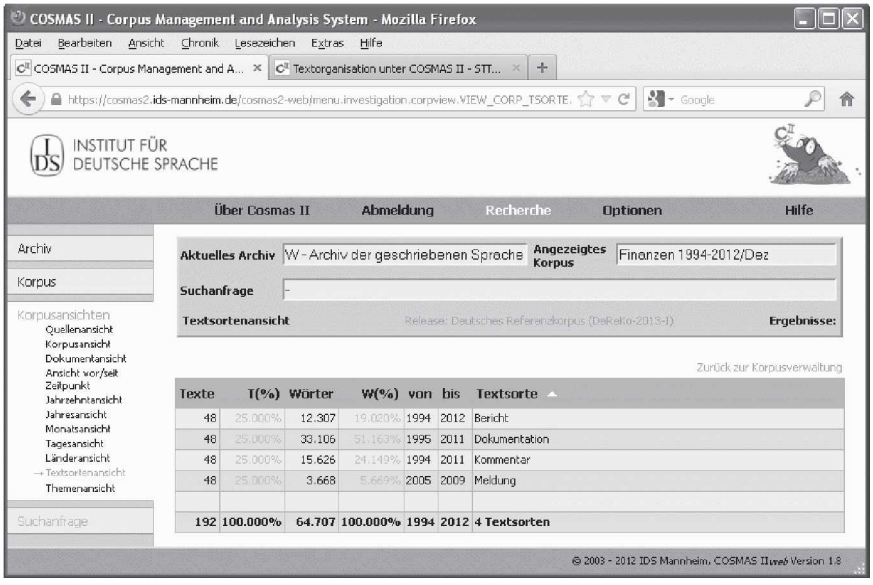

Abb. 1: Die Web-Oberfläche von COSMAS II mit Benutzerkorpus „Finanzen ...“

[1] Das Deutsche Referenzkorpus DEREKO ist mehr oder weniger ein Synonym für die IDS-Korpora, siehe www.ids-mannheim.de/kl/projekte/korpora.html sowie den Beitrag von Teubert/Belica in diesem Band.

ausgestatteten Recherche-Software einige Jahre. COSMAS II speichert in der größten Datenbank 5,5 Mrd. laufende Wortformen aus über 21 Mio. Texten oder Artikeln. Wem das nichts sagt, stelle sich eine Bücherwand mit 90.000 Büchern zu 150 Seiten bei 400 Wörtern/Seite vor.
In diesem Beitrag soll anhand einiger exemplarischer Beispiele der Umgang mit großen Textkorpora in COSMAS II veranschaulicht werden.

Stetig wachsende Korpora und Anforderungen
Die Jahr für Jahr wachsende Textmenge der DEREKO-Korpora[2] bedeutet für die an korpuslinguistischen Fragen Interessierten und das COSMAS II-Projekt ein Gewinn und eine Herausforderung zugleich. Der Gewinn wird meistens in Bereichen deutlich, wo

- Belege für seltene Sprachphänomene und Wortbildungen,
- genügend Texte für die Bildung spezialisierter Korpusquerschnitte, oder
- ausreichendes Datenmaterial für statistische Auswertungen

im Vordergrund stehen; die Herausforderung für die Nutzer besteht darin,

- aus einer manchmal unüberschaubaren Menge von Treffern möglichst viele relevante Textstellen zu erhalten und
- unerwünschte Textstellen – und möglichst nur diese – herauszufiltern;

und für die Entwickler von COSMAS II besteht sie darin,

- eine Suchmaschine zu konzipieren, die schwerlastige Suchanfragen (wie z.B. die 1,4 Mrd. Treffer von MORPH(NOUN) oder die 67.862 Komposita von *Haus* in DEREKO) in zumutbarer Zeit ausführt, wenn man nicht über eine Rechner-Farm wie die großen Internet-Dienstanbieter verfügt, und
- dem Nutzer Werkzeuge an die Hand zu geben, mit denen er durch Eingrenzen, Sortieren und Analysieren aus umfangreichen Ergebnismengen neue Erkenntnisse gewinnen kann.

Die vielen speziellen Anforderungen aus dem Gebiet der Korpuslinguistik machen es außerdem notwendig, eine Software wie COSMAS II in weiten Teilen selber zu entwickeln. Viele spezialisierte Module wurden bereits integriert: virtuelle Korpora, Zufallsreduzierung der Treffer und Korpora, Kookkurrenzanalyse, Lemmatisierung, Häufigkeitsmaße, diverse KWIC-Sortierungen,

[2] Zur Entwicklung der DEREKO-Korpora vgl. www.ids-mannheim.de/kl/projekte/korpora/archiv.html.

Korpus- und Ergebnispräsentationen etc. Parallel dazu wurden zahlreiche Optimierungsarbeiten durchgeführt, so dass z.B. aufwändige Recherchen nach Wortklassen von einer parallelen SearchEngine mit einer Geschwindigkeit von bis zu 77 Mio. Wortklassen/Sek. ausgeführt werden können. 55% aller Suchanfragen werden in weniger als einer Sekunde und 75% in weniger als vier Sekunden ausgeführt.
Auf der Basis der DeReKo-Korpora bedient der COSMAS II-Server zurzeit ca. 50 Datenbanken (Archive genannt): 24 kleinere Projektarchive (für spezielle statistische Auswertungen), 8 multilinguale Archive für das *EuroGr@mm*-Projekt, 3 morpho-syntaktisch annotierte Archive (mit insgesamt 3 Mrd. laufenden wortklassenannotierten Wortformen), ein historisches Archiv und das Hauptarchiv *W der geschriebenen Korpora* (5,4 Mrd. Wortformen), das die meisten Recherchen auf sich vereint.
Man kann allerdings nicht erwarten, dass die Zusammensetzung von Archiv *W ausgewogen* oder *repräsentativ* für die deutsche Gegenwartssprache sei. Die Intension hinter den DeReKo-Korpora ist vielmehr die Akquisition einer *Urstichprobe*[3] deutscher Schriftsprache, aus der sich möglichst viele Nutzer ihre spezifische, ausgewogene Recherchegrundlage zusammenstellen können, wie im nächsten Abschnitt ausgeführt wird.

Die Benutzerkorpora – eine virtuelle Welt
Wer also für sein Forschungsvorhaben einen oder mehrere präzise zugeschnittene Archivquerschnitte benötigt, kann sich diese mit Hilfe von neuerdings bis zu 8 Kriterien und einer Zufallsreduzierungsfunktion als *virtuelle Benutzer-Korpora* zusammenstellen (oder erstellen lassen). Es wurden über die letzten Jahre von den Nutzern ca. 2.700 dieser virtuellen Korpora angelegt.
Wie genau die Anforderungen an solche Korpuskompositionen in COSMAS II geartet sind, soll das folgende Beispiel[4] veranschaulichen:

Ziel: Bildung eines virtuellen Korpus bestehend aus *Berichten*, *Kommentaren*, etc. zum Thema *Finanzen* in den *Dezembermonaten* der letzten *20 Jahre* in *inländischen* Texten, die eine min./max. Textlänge von *50 – 1.000* Wörtern haben. Das Korpus soll in Bezug auf die Textsorten ausgewogen sein, d.h. gleich viele Texte pro Textsorte enthalten:

[3] Mehr dazu unter: www.ids-mannheim.de/kl/projekte/korpora/einsatz.html.
[4] Weitere Erläuterungen können hier nachgelesen werden: www.ids-mannheim.de/cosmas2/server/themen/korpusbildung.html.

- **Thema:**
  Wirtschaft Finanzen
- **UND Textsorte:**
  Bericht
  Dokumentation
  Kommentar
  Meldung
- **UND** durch Zufallsreduzierung
  - von jeder Textsorte gleichviele Texte
- **UND Jahresbereich:**
  Jahr = *1994 - 2012* und Monat = *Dez*
- **UND Erscheinungsland:**
  D
- **UND Textlänge:**
  50 – 1.000

Die Zusammensetzung des auf diese Weise entstandenen virtuellen Korpus kann in COSMAS II angezeigt werden. Für die Abbildung 1 wurde die Sortierung nach *Textsorten* gewählt.
Ergänzend könnten dazu analoge virtuelle Korpora für Österreich oder Deutschland oder für andere Zeitperioden (z.B. nur Januar 1994-2012) erstellt werden. Die Zerlegung in komplementäre Korpusquerschnitte bildet oft die Vorstufe zu kontrastiven Studien.

Eine *„nicht uninteressante"* Recherche
An dem nun hier folgenden Beispiel soll dargestellt werden, wie eine Vielzahl von Daten auf eine überschaubare Menge von relevanten Ergebnissen zurückgeführt werden kann:
Ein Vorhaben, dem wir kürzlich begegnet sind, befasst sich mit Adjektiv-Konstruktionen der Form: *„nicht ... un-*Adjektiv", wobei das Adjektiv mit dem Präfix *un-* gebildet wird: z.B. *„nicht unbedingt"*, *„nicht ganz unglücklich"* etc. Gewünscht ist in einem ersten Schritt eine Liste der häufigsten Adjektive, die in dieser Konstruktion erscheinen.
Dazu wählen wir das mit Wortklassen annotierte Archiv TAGGED-C (1,3 Mrd. laufende Wortformen) und finden darin:

nicht : 8 Mio.
Adjektive : 99,5 Mio.
&un- : 3 Mio.

Die Suchanfrage, die uns die gesuchten Adjektiv-Konstruktionen liefert, wobei wir einen maximalen, sinnvollen Abstand von vier Wörtern zwischen *nicht* und dem Adjektiv vorgeben wollen, lautet:

```
„nicht" /+w4,s0 (MORPH(A)/w0 &un-)
```

Dadurch erhalten wir eine Liste von immerhin noch 106.880 Adjektivkonstruktionen. Um diese Liste nach den tatsächlich vorkommenden Adjektiven sortieren zu können, erweitern wir unsere Suchanfrage mit dem #END-Operator, der das Ergebnis auf die Liste der gefundenen Adjektive einschränkt:

```
#END(„nicht" /+w4,s0 (MORPH(A)/w0 &un-))
```

Durch Sortieren (der Adjektive) nach *Wort-Types* erhalten wir eine Übersicht von immerhin noch 3.600 verschiedenen Wort-Types, von denen die häufigsten, mit einer *relativen Häufigkeit* von über 1%, in Tabelle 1 aufgelistet sind:

Tab. 1: die häufigsten *un*-Adjektive nach *nicht*

Anz. Treffer	rel. Häuf.	Wort-Type
30.398	28,441%	unbedingt
2.594	2,427%	*unmittelbar*
2.590	2,423%	*unzufrieden*
2.302	2,154%	*ungewöhnlich*
1.956	1,830%	*unverdient*
1.788	1,673%	*unumstritten*
1.507	1,410%	*unnötig*
1.304	1,220%	*unerwartet*
1.299	1,215%	*ungefährlich*
1.202	1,125%	*unmöglich*
1.079	1,010%	*unglücklich*
1.072	1,003%	*untätig*
1.070	1,001%	*unerwähnt*

Als nächstes wollen wir das mit deutlichem Abstand häufigste Adjektiv, mit 28,4% relativer Häufigkeit ist es *unbedingt*, etwas genauer untersuchen. Um die Verteilung der Abstände zwischen *nicht* und *unbedingt* herauszufinden, passen wir unsere vorherige Suchanfrage an *unbedingt* an[5] und lassen COSMAS II die Abstände zwischen den beiden Wörtern (durch Ausführen der sogenannten *statistischen KWIC-Auswertung*) zählen und anzeigen (Tabelle 2):

Treff. größe	Anz. Treffer	Anteil	Beispiel
2	29.043	95,54%	nicht unbedingt
3	769	2,53%	nicht *mehr* unbedingt
4	268	0,88%	nicht *mehr als* unbedingt
5	318	1,05%	nicht *öfter geschehen als* unbedingt
	30.398	**100,00%**	**Total**

Tab. 2: Verteilung der Abstände zwischen *nicht* und *unbedingt*

In 95,5% der Fälle folgen also beide Wörter unmittelbar aufeinander. Wir wollen trotzdem noch einen Schritt weitergehen und uns die Fälle *„nicht* X *unbedingt"* (Treffergröße 3) anschauen. Nachdem eine neue Suchanfrage nun auf einen Abstand von zwei Wörtern zwischen *nicht* und *unbedingt* eingestellt wurde[6] und die dazwischen gefundenen Wortformen nach *Wort-Types* sortiert wurden, ergibt sich folgende Auflistung (Tabelle 3):

Anz. Treffer	Anteil	nicht X unbedingt
510	67,46%	nicht *mehr* unbedingt
48	6,35%	nicht *immer* unbedingt
47	6,22%	nicht *einmal* unbedingt
28	3,70%	nicht *aber* unbedingt
20	2,65%	nicht *so* unbedingt
20	2,65%	nicht *mal* unbedingt
10	1,32%	nicht *für* unbedingt
…	…	…
769	**100,00%**	**Total**

Tab. 3: die häufigsten Wortformen in *„nicht X unbedingt"*

[5] `„nicht" /+w4,s0 unbedingt`
[6] `#NHIT(„nicht" /+w2:2,s0 unbedingt)`

Wir wollen nur noch festhalten, dass mit 67,5% relativer Häufigkeit *mehr* mit Abstand als häufigstes Wort zwischen *nicht* und *unbedingt* gefunden wurde. Im nächsten Abschnitt sollen einige weitere Nutzungen der IDS-Korpora im Hinblick auf die eingeschlagene Arbeitsweise in COSMAS II skizziert werden.

Einige weitere Beispiele von Herangehensweisen
Die hier vorgestellten Herangehensweisen spiegeln aktuelle Arbeiten oder Teile von Projekten wider und sollen die Diversitäten der Arbeiten mit großen Korpusdaten in COSMAS II aufzeigen.

Metapheranalyse: Die Sprache der Finanzen ist geprägt von der Metapher des Wassers (z.B. Geld *fließt*). Auf der Grundlage von zwei virtuellen Korpora (Texte zum Thema „Wirtschaft/Finanzen" und „Inlandpolitik" für die Zeiten 2000-2006 und 2007-heute) wird untersucht, inwiefern die Finanzkrise diese Metapher verändert hat. Zu diesem Zweck werden die Kookkurrenzprofile verschiedener Finanzbegriffe vor 2007 und ab 2007 (manuell) miteinander verglichen (Anwendung: Semesterarbeit).

Neologismen: Einige „Neulinge" lassen sich praktisch quantitativ erschließen: Es sind Wortkreationen aus der neueren Zeit, die sich bis heute etabliert haben (keine Eintagsfliegen). Schwieriger ist es, wenn ein bereits im deutschen Vokabular verankerter Begriff eine neue Bedeutung aufweist. In diesem Fall hilft z.B. wie in der vorgestellten Metapher-Studie ein Vergleich der Kookkurrenzprofile dieses Wortes in einem aktuellen und einem älteren Referenzkorpus (Steffens 2010, Anwendung: z.B. Neologismenwörterbuch).

Vorfeld: Als Teil einer grammatikalischen Untersuchung des *Vorfeldes* (Teil des Satzes vor der linken Satzklammer in Verbzweitsätzen[7]) wird aus allen Sätzen der Satzbereich vor dem ersten finiten Verb mittels einer Suchanfrage[8] extrahiert. Da diese Methode ohne grammatikalisches Wissen auskommen muss, werden auch viele unerwünschte Bereiche (die keine Vorfelder sind) mitgeliefert, die nach dem Export aus COSMAS II mit weiteren Filtern aussortiert werden. In COSMAS II werden parallel dazu die gefundenen finiten Verben mit Hilfe der *statistischen KWIC-Auswertung* nach ihrer Position im Satz sortiert, quantitativ erfasst und weiter untersucht (Anwendung: Projekt *EuroGr@mm*).

[7] Siehe: http://de.wikipedia.org/wiki/Verbzweitstellung.

[8] Die vereinfachte Suchanfrage lautet: `<sa> /+s0 #BEG(MORPH(V -INF -PCP) /s0 <se>)`.

Beobachtung der Entwicklung der Sprache: Mit drei virtuellen Korpora für die Zeitabschnitte 1905-14, 1948-57 und 1995-2005 wird die Gegenwartssprache darauf untersucht, ob z.B. Anglizismen zunehmen oder der deutsche Wortschatz bzw. die deutsche Grammatik verarmt (Anwendung: Projekt BLDS, Bericht zur Lage der Deutschen Sprache[9]).

Automatische Belegauswahl: Eine Liste von Lexikon-Einträgen wird automatisch bearbeitet, indem wie folgt vorgegangen wird: Für jedes Lemma aus der Liste wird zuerst überprüft, ob es ausreichend repräsentativ belegt ist (erscheint z.B. in mind. drei verschiedenen Quellen und Jahrgängen). In diesem Fall sollen drei verschiedene, zufällig ausgewählte Belege extrahiert werden, die nach weiteren Kriterien ausgesucht werden: aus drei versch. Quellen und Jahrgängen, nicht Teil einer Überschrift, genügend Kontext nach dem gesuchten Wort, etc. Die Automatisierung erfolgt in einem Programm, das die Anfragen und den Export der Belege aus COSMAS II durchführt (Anwendung: das *elexiko*-Projekt[10]).

Binnenmajuskeln: Das Vokabular wird auf die Entwicklung von Binnenmajuskeln (Bahn**C**ard, Inter**C**ity**E**xpress etc.) untersucht. Mit Hilfe von *regulären Ausdrücken* werden diese Wortformen aus der 42 Mio. Einträge zählenden Vokabularliste der IDS-Korpora extrahiert, um u.a. ihre Frequenz zu bestimmen (Anwendung: Magisterarbeit).

Nutzung der COSMAS II-Dienste

Ca. 30.000 Personen haben sich in den letzten 10 Jahren bei COSMAS II registriert und nutzen oder nutzten eine der beiden Benutzeroberflächen (Bodmer 2005) COSMAS II_{web} (aktuell 85%) oder COSMAS II_{win} (15%) für ihre Recherchen. Erfreulicherweise sind 81% aller Zugriffe auf die IDS-Korpora auf Nutzung von außerhalb des IDS zurückzuführen, 34% sogar auf Recherchen aus dem Ausland (Abb. 2). Die ausländischen Zugriffe wurden in Abbildung 3 noch einmal nach den wichtigsten Ländern aufgefächert.

[9] Bericht zur Lage der Deutschen Sprache: http://www.deutscheakademie.de/druckversionen/PM_Bericht zur Lage der deutschen Sprache.pdf.

[10] Vgl. www.owid.de/wb/elexiko/start.html.

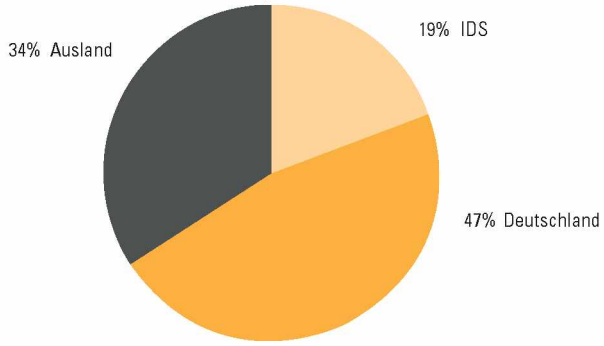

Abb. 2: Relative Verteilung der internen und externen Zugriffe

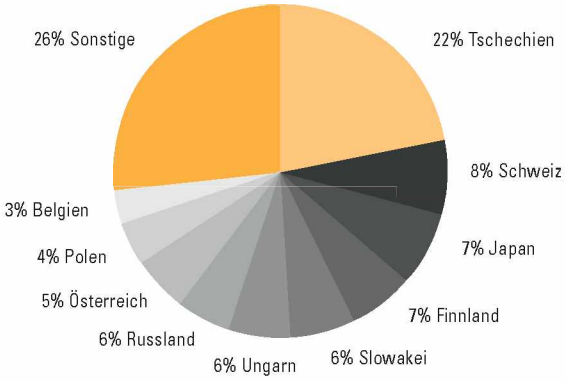

Abb. 3: Relative Verteilung der Zugriffe außerhalb Deutschlands

Seit 2003, als COSMAS II seinen Vorgänger ablöste, ist die Anzahl der Zugriffe auf die Korpora kontinuierlich von unter 50:000 pro Quartal auf 250:000 (2013/1) gestiegen. Insgesamt waren es in dieser Zeit 8,2 Mio. Zugriffe, 2,8 Mio. davon aus dem Ausland.[11] Dazu kommen gegenwärtig ca. 1 Mio. Zugriffe pro Jahr über automatisierte Prozeduren. An Tagen, an denen COSMAS II in externen Workshops eingesetzt wird, wurden schon bis zu 140 Sitzungen gleichzeitig beobachtet.

Zukunftsperspektiven

Im Kern von COSMAS II ist in den letzten Jahren eine umfangreiche Sammlung von Korpus-, Ergebnis- und Wortform-basierten Grundfunktionalitäten zur

[11] Die aktuellen Zahlen können hier eingesehen werden: http://www.ids-mannheim.de/cosmas2/projekt/nutzung/quartale.html.

Berechnung von Frequenzen, zum Sortieren und Zusammenfassen von Treffern und Korpora etc. entstanden. Auf diesem Potenzial aufbauend können in der nächsten Zeit kontinuierlich neue (korpuslinguistische) Funktionalitäten entstehen: Kombination von Ergebnis- und Korpusdaten, neue Auswertungs- und Visualisierungsmodule der Ergebnisse, Suchfilter und neue Sortierungen für das KWIC, Vergleichsfunktionalitäten für Ergebnisse und virtuelle Korpora, Erweiterungen der Suchanfragesprache und der Wortformlisten etc. Noch ganz am Anfang steht die Auswertung von Textauszeichnungen (wie: Überschriften, Fußnoten, Aufzählungen etc.), die z.B. als Filter dienen können, um Belege in unerwünschten Kontexten auszuschließen.
Der Zugang zu den IDS-Korpora über COSMAS II wird vermehrt auch in andere Projekte und vernetzte Ressourcen eingebunden werden. Zwei Szenarien lassen sich wie folgt beschreiben: Ein lexikografisches Portal (wie *OWID*[12] am IDS) veranschaulicht einen Wörterbucheintrag auf Wunsch eines Nutzers per Knopfdruck mit on-the-fly ausgewählten Korpus-Belegen.
Oder: COSMAS II agiert als Knotenpunkt in einem Netzwerk von verteilten Korpus-Kollektionen, um Suchanfragen einer spezialisierten Meta-Suchmaschine zu beantworten.
Im schnell expandierenden Bereich neuerer verteilter Datenbanktechnologien liegt allerdings der Schlüssel zu noch schnelleren Anfrage-Antwortzeiten, größeren suchbaren Textmengen und komplexer Annotationsstrukturen. Diese Ziele (unter anderem) hat sich das im Hause angesiedelte Projekt KorAP für die nächste Generation von Korpusrechercheplattformen gesteckt, die einmal COSMAS II ablösen soll.

Literatur

- **Bodmer, Franck** (2005): COSMAS II: Recherchieren in den Korpora des IDS. In: Sprachreport 3/2005, S. 2-5.
- **Steffens, Doris** (2010): Tigerentenkoalition – schon gehört? Zum neuen Wortschatz im Deutschen. In: Sprachreport 1/2010, S. 2-8.

[12] Siehe: www.owid.de/.