# Digitale Textdatenbanken im Vergleich

**Rolf Duffner (Neuchâtel)/Anton Näf (Neuchâtel)**

**Abstract**

The present paper offers a comparative presentation of the possibilities and limits of five of the most important large digital corpora on present day German which are currently publicly available (Leipzig, DWDS, COSMAS, COSMAS tagged and TIGER). The aim of this article is to put at the researcher's disposal a comprehensive survey of the possibilities for using automatic search devices in the fields of lexicology, morphology, word formation and syntax. Special attention is given to the comparison of the performance and user friendliness of the proposed search options and analysis tools (lemmatisation, cooccurrence analysis, etc.). The information given in this paper is condensed in several comparative synopses which are individually understandable to a reader who does not want to have recourse to the full text.

## 1 Einleitung

Für Forschungen im Bereich der germanistischen Linguistik besteht seit einem knappen Jahrzehnt eine völlig neue Ausgangslage. Durch die Existenz von "sehr grossen" digitalen Textdatenbanken (*very large corpora,* vgl. Belica/Steyer to come: 4) werden nämlich für Forschungen im Bereich von Lexikon und Grammatik (Morphologie, Wortbildung, Syntax) völlig neue Perspektiven eröffnet. Neben dem quantitativen Aspekt, d.h. der Möglichkeit des Zugriffs auf riesige Mengen von Sprachdaten ist aber auch auf deren Varietät bezüglich Textsorten, zeitlicher Staffelung, usw. hinzuweisen. Der wichtigste Unterschied im Vergleich zu früher ist nun aber die Möglichkeit der unentgeltlichen und dezentralisierten Nutzung von Textdatenbanken. Diese stellt einen wichtigen Schritt zur Demokratisierung der Forschung im Bereich der Sprachwissenschaft dar, verfügen doch damit der erfahrene Forscher und der Student im ersten Semester im Prinzip über gleich lange Spiesse. Beide haben nämlich Zugriff auf die einschlägigen Rohdaten und können so ihre sprachliche Intuition und ihre Hypothesen am gleichen Sprachmaterial überprüfen. Und nicht zuletzt ist es nun auch mit einem vertretbaren Aufwand möglich, den gegenwärtigen Stand der Forschung, wie er in Wörterbüchern und Grammatiken kodifiziert ist, auf den Prüfstand zu stellen. Das Interesse an solchen sehr grossen Textdatenbanken ist also ein doppeltes: Zum einen sind sie von wachsender Bedeutung im universitären Ausbildungsbetrieb, indem sie ein fast unbegrenztes Feld für Seminararbeiten, Diplomarbeiten und Dissertationen eröffnen, zum anderen werden sie für die linguistische Forschung immer unerlässlicher. Dadurch, dass man mit der Erfassung der sprachlichen Fakten des Deutschen sozusagen noch einmal ganz von vorne beginnen kann,

besteht die Hoffnung auf einen qualitativen Sprung in der Beschreibung der deutschen Gegenwartssprache.

Nachdem in der germanistischen Linguistik während über drei Jahrzehnten die theoretisch-spekulative Forschung mit dem methodischem Instrument der Elizitierung der Kompetenz des *native speaker* dominiert hatte, hat sich unterdessen angesichts des Ausbleibens von konkreten Ergebnissen eine gewisse Katerstimmung breitgemacht. Die im Gefolge der generativen Grammatik in Deutschland und anderswo entstandenen linguistischen Schulen haben sich nämlich zunehmend in Richtungskämpfe verstrickt, bei denen es oft weniger um die sprachlichen Fakten als solche als um blosse Formalisierungskonventionen ging. Bereits 1970 hatte W. Labov vor einer Sackgasse der Forschung gewarnt: *Linguists cannot continue to produce theory and data at the same time* (Labov 1972: 199). Die modernen Einsatzmöglichkeiten des Computers haben unterdessen in der Sprachwissenschaft zu einer wohltuenden Versachlichung der Debatten geführt, sodass heute wieder zunehmend die Verifizierung bzw. Falsifizierung von Hypothesen über das Funktionieren der Sprache im Zentrum steht. Jedenfalls hat sich unterdessen in weiten Kreisen die Erkenntnis durchgesetzt, dass sich auf der Grundlage von isolierten, selbstkonstruierten Beispielsätzen kaum "zeitresistente" wissenschaftliche Forschungsergebnisse erzielen lassen. Es gilt also, die Chance eines Neubeginns zu nutzen und – nun mit Hilfe der elaborierten Forschungswerkzeuge – wiederum die sprachlichen Fakten in den Blick zu nehmen. Dank leistungsfähigen Rechnern und effizienten Internetverbindungen sind seit ein paar Jahren viele Sprachen Grosskorpora für alle Interessierten weltweit zugänglich und nutzbar. Dabei ist eines der grössten Textkorpora überhaupt das Internet selbst, und die über das Internet abfragbaren Sprachdaten werden auch, meist über sogenannte Suchmaschinen wie *google* oder *alta vista* ausgewertet.[1] Allerdings erfüllt das Internet viele Standards hinsichtlich Datenbasis und Auswertungsmöglichkeiten nicht, wie sie für linguistische Forschungen unerlässlich wären. So ist etwa die Datengrundlage nicht oder nur sehr vage spezifiziert (Gesamtumfang des Korpus, Textsorten usw.), und die zur Verfügung stehenden Auswertungsprozeduren (z.B. Trefferzählung pro Zeiteinheit bei *google*) sind kaum adäquat, dies im Gegensatz zu Textdatenbanken, die ausdrücklich für sprachwissenschaftliche Bedürfnisse konzipiert wurden.

In vorliegenden Beitrag sollen nun fünf digitale Textdatenbanken zur deutschen Sprache vorgestellt und miteinander verglichen werden. Nach einer ersten summarischen Gegenüberstellung sollen Schritt für Schritt deren Einsatzmöglichkeiten im Bereich von Lexikon und Grammatik überprüft werden. Zu Beginn jedes Abschnitts präsentiert eine synoptische Tabelle überblicksartig die Hauptergebnisse des Vergleichs, die dann in einem Fliesstext präzisiert und kommentiert werden. Auswahlkriterium für die zu vergleichenden Korpora war in erster Linie deren kostenlose und ortsunabhängige Nutzung. Ansonsten sind die Datenbanken so unterschiedlich wie die Fragestellungen, für die sie eingesetzt werden können.

Wir möchten abschliessend ausdrücklich betonen, dass wir von Haus aus nicht Computerlinguisten, sondern Sprachwissenschaftler sind und dass hier deshalb nicht die technische Seite

[1] Bei der Erarbeitung des Variantenwörterbuchs des Deutschen wurden sehr umfangreiche Internetrecherchen durchgeführt, vgl. dazu Bickel (2000).

der Programme im Vordergrund steht. Vielmehr nehmen wir hier ganz bewusst die Perspektive eines Benutzers ein, der mit vernünftigem Lernaufwand möglichst schnell zu sprachwissenschaftlich relevanten Ergebnissen gelangen will.

## 2 Charakteristika der fünf Textdatenbanken

Im Folgenden sollen nun fünf Textdatenbanken und deren Such- und Analysewerkzeuge summarisch vorgestellt und miteinander verglichen werden,[2] nämlich die Wortschatz-Datenbank der Universität Leipzig (im Folgenden: *Leipzig*), das Digitale Wörterbuch der deutschen Sprache des 20. Jahrhunderts (*DWDS*), das Korpus des Instituts für deutsche Sprache (*Cosmas*),[3] das morphologisch annotierte Teilkorpus von *Cosmas* (*Cosmas tagged*) sowie das Sampler-Corpus zu TIGERsearch (*Tiger*).

| | Leipzig | DWDS | Cosmas | Cosmas tagged | Tiger |
|---|---|---|---|---|---|
| Technische Angaben | | | | | |
| *Adresse* | wortschatz.uni-leipzig.de | www.dwds.de | www.ids-mannheim.de | www.ids-mannheim.de | www.tigersearch.de |
| *Quelle* | Uni Leipzig | BBAW | IDS | IDS | DFG |
| *Download* | nein | nein | ja | ja | ja |
| *Arbeit online* | ja | ja | ja | ja | nein |
| *Betriebssystem* | diverse | diverse | diverse | diverse | OSx/Win98/NT/Linux |
| *Grösse (Wortzahl)* | 500 Mio | 102 Mio | 940 Mio öff. zugänglich | 30 Mio | 0.7 Mio |
| *Teilkorpora anwählbar* | nein | ja | ja, auch virtuelle Korpora | ja, auch virtuelle Korpora | nein |
| *Textsorten* | v.a. Zeitungen | 4 Domänen | v.a. Tageszeitungen | v.a. Spiegel und Mannheimer Morgen | Frankfurter Rundschau |
| *Gesprochene Sprache* | nein | nein | z. T. vorhanden | z. T. vorhanden | nein |
| *Zeit* | nach 1989 | 1900 - 2000 (in Dekaden) | v.a. 1990 - 2000 | v.a. 1990 - 2000 | nach 1990 |
| *kostenpflichtig* | nein | nein | nein | nein | nein |
| Hauptsuchoptionen | | | | | |
| *Lemmasuche* | nein | ja | ja | ja | nein |
| *Kookkurrenzanalyse* | ja (eingeschränkt) | ja | ja | ja | nein |
| *Annotierung Morphologie (tagging)* | nein | ja | nein | ja | ja |
| *Annotierung Syntax (parsing)* | nein | nein | nein | nein | ja |

Tabelle 1: Charakteristika der fünf digitalen Textdatenbanken

[2] Internet-Adressen der Korpora vgl. Tabelle 1.

[3] COSMAS ist ein Akronym für Corpus Search, Management and Analysis System. Die Version 1 von COSMAS II ist seit 1995 im Einsatz (ab 2002 Version 3, aktuell Version 3.6).

Die beiden *Cosmas*-Datenbanken wurden vom Institut für deutsche Sprache (IDS) in Mannheim erstellt und werden dauernd erweitert und – nicht zuletzt durch eine ständige Verbesserung der Abfrage- und Darstellungsmöglichkeiten – für die Forschung aufbereitet. Das *DWDS* ist der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW) angegliedert, während *Leipzig* von der Abteilung Automatische Sprachverarbeitung des Instituts für Informatik der Universität Leipzig erstellt wurde. Das Demonstrationskorpus *Tiger* und das dazugehörige Suchwerkzeug *TIGERsearch* wurde mit Mitteln der deutschen Forschungsgemeinschaft (DFG) an den Universitäten Saarbrücken, Stuttgart und Potsdam entwickelt.

Alle fünf Textdatenbanken sind kostenlos sowie ortsunabhängig nutzbar. *Leipzig* und *DWDS* bieten darüber hinaus den Vorteil, dass sie online genutzt werden können und dass somit keine grossen Datenmengen auf der eigenen Festplatte untergebracht werden müssen. *Cosmas* kann ebenfalls online genutzt werden, vorgängig muss allerdings ein Anwendungsprogramm (Windows-Client) installiert werden (1.2 MB). Demgegenüber kann man mit *Tiger* zwar netzunabhängig arbeiten, dafür muss aber das gesamte Korpus auf der Festplatte gespeichert werden. Sämtliche Datenbanken laufen unterdessen auf allen gängigen Betriebssystemen. Sowohl *Cosmas* als auch *DWDS* verfügen (teilweise aus Gründen des Urheberschutzes) zusätzlich noch über nicht öffentlich zugängliche Teilkorpora.

Das mit Abstand grösste Textkorpus ist zur Zeit (Stand 2005) *Cosmas* mit 940 Mio öffentlich zugänglichen Textwörtern, gefolgt von *Leipzig* (500 Mio), *DWDS* (102 Mio), *Cosmas tagged* (30 Mio) und – mit grossem Abstand – vom morphologisch und syntaktisch annotierten *Tiger*-Korpus (0,7 Mio).

Bei *Leipzig* ist – im Gegensatz zu den andern Korpora – kaum etwas über die Auswahl der Quellen (offenbar vor allem Zeitungstexte) in Erfahrung zu bringen. Am ausgewogensten bezüglich der Wahl der Textsorten (mit den vier einzeln anwählbaren Domänen *Zeitung, Belletristik, Wissenschaft, Gebrauchsliteratur*) ist *DWDS*, während *Cosmas* in erster Linie Zeitungen und Zeitschriften zur Verfügung stellt. Allerdings erlaubt *Cosmas* den gezielten Zugriff auf eine grosse Zahl einzelner Teilkorpora (verschiedene Tageszeitungen, literarische Werke, usw.). *Leipzig* und *Cosmas* repräsentieren schwergewichtig die Sprache ab zirka 1990, während das diachronisch konzipierte *DWDS* auf gleichmässige Weise die Dekaden des 20. Jahrhunderts abdeckt.

Wie steht es nun mit der Benutzerfreundlichkeit? Am leichtesten und mehr oder weniger intuitiv zugänglich ist zweifellos *Leipzig*, während man von *Tiger* nur nach einem Studium des Benutzerhandbuchs wirklich profitieren kann. *Cosmas* stellt von den nicht annotierten Korpora die raffiniertesten Analysemethoden zur Verfügung; das Ausschöpfen aller Nutzungsmöglichkeiten setzt allerdings einiges an Übung voraus.

## 3 Abfragemöglichkeiten im Bereich des Lexikons

Die untenstehende Tabelle 2 fasst die Abfragemöglichkeiten der fünf Textdatenbanken im Bereich der Lexikologie zusammen.

| | Leipzig | DWDS | Cosmas | Cosmas tagged | Tiger |
|---|---|---|---|---|---|
| Abfrage | | | | | |
| *Wort (im Sinne von Wortform)* | ja | ja | ja | ja | ja |
| *Wortbestandteile* | ja | ja | ja | ja | bedingt |
| *Lemma* | nein | als Voreinstellung | ja | ja | nein |
| *Boolsche Op. (AND, OR, NOT)* | nein | ja | ja | ja | ja |
| *Kombination von Suchbegriffen* | nein | ja | ja | ja | bedingt |
| *Suche mit definiertem Abstand* | nein | ja | ja | ja | nein |
| *Kookkurrenzanalyse* | nach Frequenz | ja | ja | ja | nein |
| *Funktionswörter ignorierbar* | als Voreinstellung | nein | ja | ja | – |
| Darstellung | | | | | |
| *Anzahl Treffer* | ja | ja | ja | ja | ja |
| *Wortformenliste* | ja, aber unhandlich | nein | ja | ja | nein |
| *KWIC (Konkordanz)* | bedingt | ja | ja | ja | nein |
| *KWIC verschieden sortierbar* | nein | nein | ja | ja | nein |
| *Maximaler Kotext* | Satz | mehrere Sätze | mehrere Sätze | mehrere Sätze | Satz |
| *Kookkurrenzliste* | bloss Frequenzliste | ja | ja | ja | nein |
| *Ergebnisse exportierbar* | bedingt | ja | ja | ja | ja |

Tabelle 2: Abfragemöglichkeiten im Bereich des Lexikons

Ein bestimmtes Wort bzw. eine bestimmte Wortform suchen kann jedes der hier beschriebenen Suchprogramme, so wie im Prinzip auch jede Internet-Suchmaschine. Für spezifischere Forschungen im Bereich der Lexikologie des heutigen Deutsch bietet jedoch eindeutig *Cosmas* die breiteste Palette von Suchwerkzeugen. Um beim Elementarsten zu beginnen: Sowohl *Cosmas* als auch *DWDS* (hier als Voreinstellung eingerichtet) können einzelne Wortformen einer Grundform (Lemma) zuordnen. Mit anderen Worten: diese beiden Programme können im Prinzip die Wortformen *Zug, Zuges, Züge* usw. dem Lemma *ZUG* zuweisen, und sie liefern umgekehrt unter ZUG alle im Korpus vorkommenden Wortformen. Dies ist nicht zuletzt bei suppletiven Paradigmen nützlich, beispielsweise bei der Zuordnung von Formen wie *ist, seid, war, bist* und *gewesen* zur Zitierform SEIN. Ein grosser Nachteil von *Leipzig* ist das Fehlen eben dieser Suchoption.

*Leipzig* ist den beiden anderen Suchwerkzeugen auch insofern unterlegen, als es keine Suche mit Boolschen Operatoren erlaubt (AND, OR, NOT), ein für linguistische Zwecke entscheidender Mangel.

Sowohl *Cosmas* als auch *DWDS* erlauben die Kombination sowohl von Suchwörtern als auch von Suchoperatoren, und zwar mit zwei oder mehr Leerstellen (binäre und multiple Kombinationen). Als Rahmeneinheit der Suche dient dabei der Satz. So kann man etwa bei *Cosmas* nicht nur nach zwei oder mehr Wortformen suchen, sondern auch nach einer Wortform und einem Lemma oder nach zwei alternativen Wortformen und einem Lemma, usw. Dabei können der Abstand (Anzahl Wörter) und die Richtung der Suchbegriffe (Suche nach links oder nach rechts oder beides zugleich) durch Voreinstellung definiert werden.

Das Herzstück jeder linguistischen Textkorpusanalyse ist indes die statistische Kookkurrenzanalyse. Diese dient der Vorstrukturierung sprachlicher Massendaten.

Durch die Kookkurrenzanalyse werden distributionelle Eigenschaften von Wörtern und Zeichenketten erfasst, die im Vergleich zu ihrem Gesamtvorkommen statistisch überproportional häufig in der Umgebung anderer Zeichenkonfigurationen vorkommen. Die Kookkurrenzanalyse erfasst diese Zeichenketten und ordnet sie in hierarchischen Kookkurrenzclustern an (vgl. Belica/Steyer to come: 8ff.). Programme zur Kookkurrenzanalyse sind in *Cosmas, DWDS* und *Leipzig* bereits integriert. Allerdings muss gleich dazu gesagt werden, dass die in *DWDS* und besonders in *Leipzig* implementierten Möglichkeiten der Kookkurrenzanalyse im Verlgleich zu jenen von *Cosmas* eher bescheiden sind. Bei *Leipzig* etwa ist die Kookkurrenzanalyse auf zwei fest vorgegebene Optionen beschränkt: *Suche Adjektiv zu Substantiv* und *Suche Verb zu Substantiv* (der jeweils umgekehrte Suchbefehl ist beispielsweise nicht möglich). Eine solche Abfrage kann gewiss interessante globale Ergebnisse liefern, ist aber für linguistische Fragestellungen bedeutend weniger relevant als das in *Cosmas* integrierte Analyseprogramm mit seinen zahlreichen verstellbaren Parametern für gezieltere Suchabfragen.

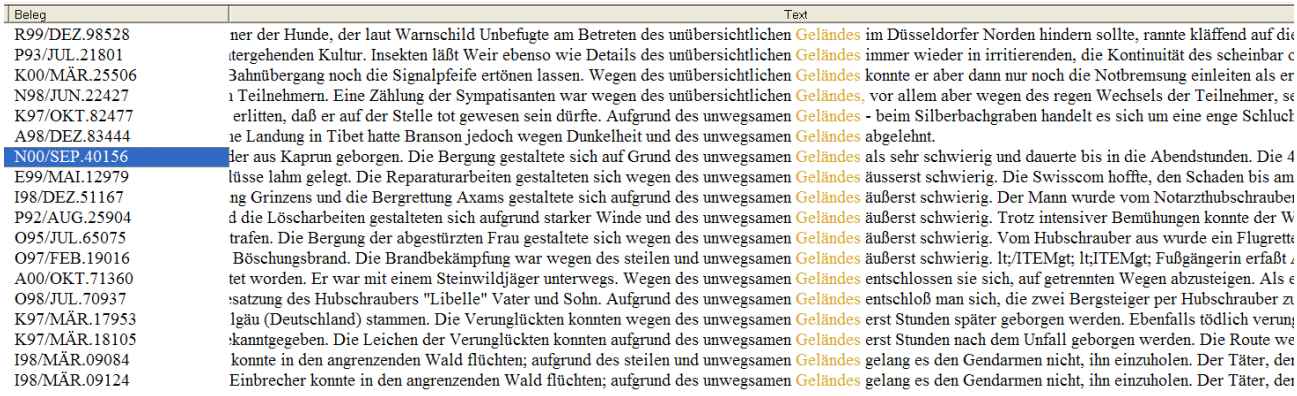
Einer dieser einstellbaren Parameter bei *Cosmas* ist die Option *Funktionswörter ignorieren*. Unter der Bezeichnung Funktionswörter werden geschlossenene Klassen von "Kleinwörtern" zusammengefasst (Artikelwörter, Pronomen, Präpositionen, Konjunktionen, usw.). Da diese in der Regel eine sehr hohe Kofrequenz mit den Hauptwortarten (Nomen, Verb, Adjektiv) aufweisen, erweist es sich oft als sinnvoll (und darüberhinaus zeitsparend), diese von der Analyse auszuschliessen. Damit wird zum einen die Kookkurrenzanalyse beschleunigt und zum andern das Lesen und Interpretieren des Kookkurrenzprofils erleichtert. Im Gegensatz dazu sind etwa bei *Leipzig* die Funktionswörter zum vornherein von der Analyse ausgeschlossen. Dies kann aber zu schwerwiegenden Fehlinterpretationen führen, etwa im Falle der Verbzusätze in Klammerstellung. Bei einer Suche nach den Belegen von *einstellen* oder *beilegen* beispielsweise werden nämlich die in Klammerstellung befindlichen Wortbestandteile *ein* oder *bei* fälschlicherweise den Funktionswörtern zugeschlagen und damit von der Analyse ausgeschieden (für das Suchprogramm liegen hier Belege für die einfachen Verben *stellen* und *legen* vor).

Neben dem Angebot an Abfrageoptionen tragen auch die verschiedenen Möglichkeiten der graphischen Darstellung von Ergebnissen und Zwischenresultaten entscheidend zur *Performance* eines Suchwerkzeugs bei. Im Bereich Lexikon sind das im wesentlichen Wortformenlisten, Beleglisten und Kookkurrenzlisten. Für etwas pointiertere linguistische Fragestellungen verfügt auch hier nur *Cosmas* über ein zufriedenstellendes Instrumentarium. Im Gegensatz zu *Cosmas* ist bei *DWDS* die Option Wortformenliste nicht vorgesehen, und diejenige von *Leipzig* ist nur schwer zu handhaben. Die Wortformenliste bei *Cosmas* – etwas unglücklich als *Suchbegriff-Expansionsliste* bezeichnet – enthält mehrere für die Forschung relevante Suchfunktionen. So liefert sie bei einer Lemmasuche die im Korpus gefundenen *types* samt der jeweiligen Anzahl *tokens* der einzelnen Flexionsformen. In diesen Wortformenlisten wird überdies zwischen Gross- und Kleinschreibung der *types* unterschieden, was unter anderem für orthographische, aber auch für syntaktische Fragen (z. B. Häufigkeit der Besetzung der Erstposition im Satz) von Interesse ist.

Im weiteren ist bei *Cosmas* sehr nützlich, dass einzelne Wortformen jeweils nach Wunsch von der anschliessenden Suche bzw. Analyse ausgeschlossen werden können. Die Wortformenliste kann dabei sowohl alphabetisch als auch nach Frequenz geordnet werden. Solche Listen sind etwa bei Untersuchungen zur Komposition von grossem Nutzen.

Die für den Belegausdruck wichtigste – nicht zuletzt weil platzsparende – Form der Darstellung von Belegen ist jene in Konkordanzform (Suchwort in der Mitte der Zeile in Fettdruck samt einem Vor- und Nachlauf von einigen Wörtern). Eine solche Art der Darstellung von Kurzbelegen – meist KWIC (Key Word In Context) genannt – ist heute weitgehend zum Standard geworden und dementsprechend bei den beiden elaborierteren Suchprogrammen (*Cosmas* und *DWDS*) direkt implementiert. *Cosmas* bietet darüber hinaus die Möglichkeit, die KWIC-Belege nach verschiedenen Gesichtspunkten zu sortieren, nämlich nach den Parametern "Quelle", "chronologisch" oder "alphabetisch" (mit den Unteroptionen alphabetisch nach dem Keyword und/oder dem ersten, zweiten oder dritten Kontextwort links oder rechts). Besonders dank der letztgenannten Möglichkeit werden bestimmte syntaktische Muster (z.B. Adjektiv-Nomen- oder Präposition-Nomen-Verbindungen) sozusagen auf einen Blick sichtbar. So geht etwa aus dem KWIC-Ausschnitt von Fig. 1 unmittelbar hervor, dass das Nomen *Gelände* vor allem mit *unwegsam* und *unübersichtlich* kombiniert auftritt, häufig als Teil einer Präpositionalgruppe mit *wegen/aufgrund* und innerhalb typischer verbaler Konstruktionen wie *(nicht) gelingen* und *sich (äusserst) schwierig gestalten* sind, was folgenden prototypischen Satz ergibt: *Die Rettungsarbeiten gestalteten sich aufgrund des unwegsamen Geländes äusserst schwierig*.

ISSN 1615-3014

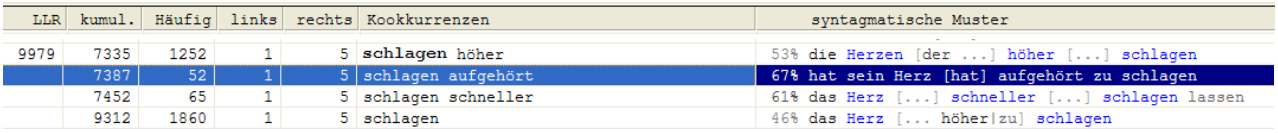


| Beleg | Text |
|---|---|
| R99/DEZ.98528 | ner der Hunde, der laut Warnschild Unbefugte am Betreten des unübersichtlichen Geländes im Düsseldorfer Norden hindern sollte, rannte kläffend auf die |
| P93/JUL.21801 | ıtergehenden Kultur. Insekten läßt Weir ebenso wie Details des unübersichtlichen Geländes immer wieder in irritierenden, die Kontinuität des scheinbar c |
| K00/MÄR.25506 | Bahnübergang noch die Signalpfeife ertönen lassen. Wegen des unübersichtlichen Geländes konnte er aber dann nur noch die Notbremsung einleiten als er |
| N98/JUN.22427 | ı Teilnehmern. Eine Zählung der Sympatisanten war wegen des unübersichtlichen Geländes, vor allem aber wegen des regen Wechsels der Teilnehmer, se |
| K97/OKT.82477 | erlitten, daß er auf der Stelle tot gewesen sein dürfte. Aufgrund des unwegsamen Geländes - beim Silberbachgraben handelt es sich um eine enge Schluch |
| A98/DEZ.83444 | ıe Landung in Tibet hatte Branson jedoch wegen Dunkelheit und des unwegsamen Geländes abgelehnt. |
| N00/SEP.40156 | der aus Kaprun geborgen. Die Bergung gestaltete sich auf Grund des unwegsamen Geländes als sehr schwierig und dauerte bis in die Abendstunden. Die 4 |
| E99/MAI.12979 | lüsse lahm gelegt. Die Reparaturarbeiten gestalteten sich wegen des unwegsamen Geländes äusserst schwierig. Die Swisscom hoffte, den Schaden bis am |
| I98/DEZ.51167 | ng Grinzens und die Bergrettung Axams gestaltete sich aufgrund des unwegsamen Geländes äußerst schwierig. Der Mann wurde vom Notarzthubschrauber |
| P92/AUG.25904 | d die Löscharbeiten gestalteten sich aufgrund starker Winde und des unwegsamen Geländes äußerst schwierig. Trotz intensiver Bemühungen konnte der W |
| O95/JUL.65075 | trafen. Die Bergung der abgestürzten Frau gestaltete sich wegen des unwegsamen Geländes äußerst schwierig. Vom Hubschrauber aus wurde ein Flugrette |
| O97/FEB.19016 | Böschungsbrand. Die Brandbekämpfung war wegen des steilen und unwegsamen Geländes äußerst schwierig. lt;/ITEMgt; lt;ITEMgt; Fußgängerin erfaßt / |
| A00/OKT.71360 | tet worden. Er war mit einem Steinwildjäger unterwegs. Wegen des unwegsamen Geländes entschlossen sie sich, auf getrennten Wegen abzusteigen. Als e |
| O98/JUL.70937 | satzung des Hubschraubers "Libelle" Vater und Sohn. Aufgrund des unwegsamen Geländes entschloß man sich, die zwei Bergsteiger per Hubschrauber zu |
| K97/MÄR.17953 | lgäu (Deutschland) stammen. Die Verunglückten konnten wegen des unwegsamen Geländes erst Stunden später geborgen werden. Ebenfalls tödlich verun |
| K97/MÄR.18105 | kannt gegeben. Die Leichen der Verunglückten konnten aufgrund des unwegsamen Geländes erst Stunden nach dem Unfall geborgen werden. Die Route we |
| I98/MÄR.09084 | konnte in den angrenzenden Wald flüchten; aufgrund des steilen und unwegsamen Geländes gelang es den Gendarmen nicht, ihn einzuholen. Der Täter, der |
| I98/MÄR.09124 | Einbrecher konnte in den angrenzenden Wald flüchten; aufgrund des unwegsamen Geländes gelang es den Gendarmen nicht, ihn einzuholen. Der Täter, der |

Fig. 1

All diesen Komfort bieten *DWDS* und insbesondere *Leipzig* nicht. Bei *Leipzig* erscheint zwar eine (durch Zufallsauswahl generierte?) Liste von Belegen in Form ganzer Sätze, bei der immerhin das Suchwort durch Fettdruck hervorgehoben ist. Aber "mit dem Finger herunterfahrend" ins Auge springende "Wortnester" und Strukturen auf einen Schlag erkennen kann man hier nicht. Der maximal abfragbare Kotext besteht bei *Leipzig* aus einem Satz, was für die linguistische Forschung (etwa semantische Disambiguierungen) manchmal zu knapp ist. *DWDS* und *Cosmas* liefern in der Regel – als Wahlmöglichkeit neben KWIC und dem Kontext eines Satzes – mehrere Sätze bzw. einen ganzen Abschnitt.

Der Hauptvorteil von *Leipzig* ist wohl, dass der Nutzer schnell und ohne grossen Lernaufwand zu einer recht übersichtlichen Kookkurrenzliste kommt. Allerdings kann er dann an dieser keine weiteren Operationen mehr vornehmen. Auch *DWDS* liefert eine solche Liste mit diversen statistischen Angaben, aber auch hier können zum Beispiel die Belegsätze für die einzelnen Kookkurrenten nicht einfach herbeigeklickt werden. Diesbezüglich weit überlegen ist die Darstellung der Kookkurrenzen bei *Cosmas*, da hier nicht bloss (wie bei *DWDS*) binäre Kookkurrenzen, sondern auch elaborierte multiple Kookkurrenzcluster zur Verfügung gestellt werden. Beispiel: Zu *Herz* erscheint nicht einfach nur das Kollokat *schlagen*, sondern darüber hinaus speziellere multiple Cluster wie *Herz+schlagen+höher* oder *Herz+schlagen+aufgehört*, usw. Dabei wird für jede Kombination die Frequenz und für die Hauptkookkurrenz *schlagen* der LLR-Wert (*logarithmic likelihood ratio*), ein Mass für die "Exklusivität" der Beziehung zwischen zwei Argumenten, angegeben, vgl. Fig. 2.



| LLR | kumul. | Häufig | links | rechts | Kookkurrenzen | syntagmatische Muster |
|---|---|---|---|---|---|---|
| 9979 | 7335 | 1252 | 1 | 5 | **schlagen** höher | 53% die Herzen [der ...] höher [...] schlagen |
| | 7387 | 52 | 1 | 5 | schlagen aufgehört | 67% hat sein Herz [hat] aufgehört zu schlagen |
| | 7452 | 65 | 1 | 5 | schlagen schneller | 61% das Herz [...] schneller [...] schlagen lassen |
| | 9312 | 1860 | 1 | 5 | schlagen | 46% das Herz [... höher\|zu] schlagen |

Fig. 2

Einzelne Textdatenbanken bieten darüberhinaus noch weitere Möglichkeiten, die im Bereich Lexikon von Interesse sein können. *Leipzig* etwa liefert auf Wunsch eine graphische Darstellung der Kookkurrenzen (deren Interpretation allerdings nicht gerade auf der Hand liegt) und dazu verschiedene Angaben aus Wörterbüchern. Auch *DWDS* hält solche lexikographische Zusatzinformationen bereit.

Was hingegen das Herz jedes Lexikologen – und wohl noch stärker dasjenige jedes Lexikographen – höher schlagen liesse, ist bislang allerdings noch kein automatisches Suchprogramm zu leisten im Stande: die Zuordnung der Belege zu den einzelnen Unterbedeutungen eines Lexems.[4] Um beim Verb *einstellen* zu bleiben: Bei einem Kollokator wie *(Straf)verfahren* "weiss" jeder kompetente Sprecher sofort, dass es sich hier um die Unterbedeutung ‚nicht fortsetzen' handelt, bei *Personal* um ‚anstellen', bei *Lautstärke* um ‚regulieren' usw. Auch wenn eine solche im höchsten Masse praxisrelevante Anwendung vorläufig noch ausserhalb der technischen Möglichkeiten der digitalen Datenverarbeitungsprogramme liegt, sind diese bereits heute für die Wörterbuchherstellung von allergrösstem Nutzen. Durch das automatische Durchforsten immenser Textmengen und der benutzerfreundlichen Aufbereitung und Darbietung der Ergebnisse in Form von Kookkurrenzlisten wird die Arbeit des Lexikographen ungemein erleichtert. Dieser kann sich nun nämlich – von der mühsamen und zeitaufwändigen Belegsuche befreit – ganz der Interpretation und optimalen Verwertung des Belegmaterials widmen.

## 4 Abfragemöglichkeiten im Bereich der Morphologie

Die untenstehende Tabelle 3 fasst die Abfragemöglichkeiten der fünf Textdatenbanken im Bereich der Morphologie und Wortbildung zusammen.

Synopse der Abfragemöglichkeiten im Bereich der Morphologie

| | Leipzig | DWDS | Cosmas | Cosmas tagged | Tiger |
|---|---|---|---|---|---|
| *Mechanische Suche nach Wortbestandteilen (*)* | ja | ja | ja | ja | ja |
| *Suche nach Wortklasse* | nein | ja | nein | ja | ja |
| *Suche nach Kategorien (z.B. Numerus, Kasus, Tempus)* | nein | nein | nein | ja | ja |
| *"Intelligente" Suche nach Wortbestandteilen* | nein | jein | nein | ja | ja |

Tabelle 3: Abfragemögichkeit im Bereich der Morphologie (inkl. Wortbildung)

Das morphologisch annotierte Teilkorpus von *Cosmas*, hier als *Cosmas tagged* abgekürzt, ist speziell bei Fragen im Bereich Morphologie und Wortbildung von grossem Nutzen. Auch *DWDS* und das *Tiger*korpus wurden mit mit Hilfe eines *Taggers* linguistisch annotiert. Die Eingabesprache bei *DWDS* im Bereich Morphologie muss allerdings als wenig benutzerfreundlich bezeichnet werden. Auch *Tiger* ist konzeptbedingt für morphologische Fragestellungen weniger geeignet – einmal ganz absehen vom geringen Umfang des Demonstrationskorpus.

Alle fünf Textdatenbanken lassen eine mechanische Suche nach Wortbestandteilen zu, meist mit Hilfe des Asterisks (*) als Platzhaltersymbol. So liefern etwa die Suchwerkzeuge auf die

[4] Allerdings muss hier gleich hinzugefügt werden, dass das der Datenbank *Cosmas* zugrunde liegende Konzept nicht von vorgegebenen semantischen Unterscheidungen ausgeht, sondern die Daten der Forschung uninterpretiert – aber mit statistischen Methoden aufbereitet – zur Verfügung stellen will.

Anfrage **bar* eine grosse Zahl von Adjektiven (*wunderbar, sichtbar, spürbar,* usw.), aber natürlich auch Substantive wie *Nachbar, Escobar* oder gar *Cüpli-Bar*.

*Cosmas tagged* und *DWDS* verfügen – im Gegensatz zu *Leipzig* und *Cosmas* – über die Option, nicht bloss konkrete Lexeme, sondern auch ganze Wortklassen (Nomen, Verb, Präposition, usw.) anzupeilen. Dank diesem *Tool* – bei *Cosmas tagged* morphosyntaktischer Assistent genannt – kann in der Anfrage eine Wortform bzw. ein Wortbestandteil mit einer Wortklasse als zweitem Argument (z.B. *Adjektiv*) kombiniert werden. Mit einer solchen "intelligenten" Kombinationssuche findet *Cosmas tagged* im Korpus etwa alle Wortformen auf *-bar*, die zugleich Adjektive sind. Bei *DWDS* sind ähnliche Abfragen möglich. Allerdings kennt dieses Suchwerkzeug lediglich den Parameter *Wortklasse*, darüberhinaus jedoch keine grammatischen Kategorien wie Numerus oder Kasus.

Ausgehend vom Beispielsatz

(1) *Charles […] war der begehrteste Junggeselle der Welt*. (Die Presse, 11.12.1992)

kann man beispielsweise mit *Cosmas tagged* alle anderen Syntagmen mit einem Adjektiv im Superlativ und *Welt* in der Rolle eines Genitivattributs suchen; als Resultate erscheinen dann zum Beispiel: *der grösste Markt der Welt*, *die beste Band der Welt,* usw. Derartige "intelligente" Kombinationssuchen von lexikalischen und grammatischen Elementen sind für die linguistische Forschung von hohem Interesse.

Automatisch generierte morphologische Annotierungen sind bislang allerdings noch nicht völlig fehlerfrei, besonders was die Zuweisung von Kategorien wie Kasus und Tempus betrifft. Zudem ist jede Annotierung abhängig von der dahinterstehenden Grammatiktheorie. Die Suchsoftware-Konzeptoren sind dabei mit zahlreichen Fragen konfrontiert, zum Beispiel: Soll eine Wortklasse *Adverb* angesetzt werden und soll diese Klasse gegebenenfalls noch feiner in Untergruppen differenziert werden? Und wenn ja, nach syntaktischen und/oder semantischen Kriterien? Im grossen und ganzen steht hinter den Annotierungen von *Cosmas tagged* und *DWDS* die Wortartenlehre der traditionellen Schulgrammatik. Eine solche – konservative – Einteilung im Bereich der unflektierbaren Wörter erscheint indes trotz deren manifester Schwächen als vertretbar, dies deswegen, weil durch spezifischere und zum Teil schulabhängige Klassifizierungen das sprachliche Rohmaterial durch terminologieimmanente Deutungen vorinterpretiert würde. Die Entwickler von *DWDS* haben laut eigenen Angaben auf das Stuttgart-Tübinger Tagset zurückgegriffen, diejenigen von *Cosmas tagged* auf das MECOLB-Minimal Tagset.

## 5 Abfragemöglichkeiten im Bereich der Syntax

Die Erforschung von syntaktischen Fragestellungen setzt im Idealfall ein Korpus voraus, das sowohl mit einem *Tagger* (Zuweisung von Wortklassen und grammatischen Kategorien) als auch mit einem *Parser* (Zuweisung von syntaktischen Funktionen wie z. B. Satzgliedrollen) "vorbehandelt" wurde. Da eine derartige grammatische Annotierung in *Leipzig* und *Cosmas* nicht vorgenommen wurde, sind diese beiden Suchprogramme im Prinzip für syntaktische Forschung nur bedingt geeignet. Immerhin bietet *Cosmas* im Vergleich zu allen andern

Programmen den unschätzbaren Vorteil, dass die Zwischenergebnisse einer Suchanfrage weiter behandelt werden können. So können etwa KWIC-Belege einzeln angewählt und in einem virtuellen Korpus abgespeichert werden, welches dann die Grundlage für weitere Datenmanipulationen bilden kann. Bei derartigen Ergebnismanipulationen kann der Benutzer seine linguistischen Kenntnisse zwischenschalten und so dem Suchwerkzeug auf intelligente Weise "auf die Sprünge helfen". Diesen Vorgang des schrittweisen Sich-Annäherns an die anvisierte Suchfrage wollen wir im Folgenden als *Anfragezuspitzung* beizeichnen (Habert spricht im gleichen Sinne von *ajustements successifs*). Das *DWDS*-Korpus ist zwar morphologisch annotiert, lässt aber keine dem *Cosmas* vergleichbaren Möglichkeiten der Anfragezuspitzung zu. *Cosmas tagged* seinerseits kumuliert den Vorteil der morphologischen Annotierung mit der Möglichkeit des Veränderns und Weiterverwendens von Zwischenergebnissen. Insgesamt gilt aber, dass bloss *Tiger*, dessen Korpus mit einem *Parser* – und offenbar viel Handkorrektur – für im engeren Sinne syntaktische Fragestellungen konzipiert wurde, eine breite Palette von Instrumenten zur Erforschung der deutschen Syntax bietet.

Synopse der Abfragemöglichkeiten im Bereich der Syntax

| | Leipzig | DWDS | Cosmas | Cosmas tagged | Tiger |
|---|---|---|---|---|---|
| *Funktionale Trennung von homonymen Wortkörpern* (als) | nein | ja | nein, z.T. durch Anfragezuspitzung möglich | ja | ja |
| *Funktionale Trennung Flexionsmorphemen* (schön-er) | nein | nein | nein, z.T. durch Anfragezuspitzung möglich | ja | ja |
| *Funkionale Trennung von Nominalgruppen* (*NG im Genitiv:* des Schutzes) | nein | nein | nein | nein, aber durch kreative Suchanfragen z.T. möglich | ja |
| *Valenz: Satzbaupläne* (SUBJ+AKK+DAT) | nein | nein | nein | nein | ja |
| *Funktionale Trennung von identischen Wortketten* (Bin ich ein Versager ?/!/,) | nein | nein | nein, z.T. durch Anfragezuspitzung möglich | nein, z.T. durch Anfragezuspitzung möglich | nein, z.T. durch Anfragezuspitzung möglich |

Tabelle 4: Abfragemöglichkeiten im Bereich der Syntax

Anhand von fünf typischen syntaktischen Fragestellungen (vgl. Tabelle 4) soll nun die Eignung der einzelnen Textdatenbanken, vor allem *Cosmas (tagged)* und *Tiger*, für die Syntaxforschung geprüft und veranschaulicht werden.

Erstes Beispiel: Funktionale Trennung von homonymen Wortkörpern

Jeder kompetente Sprecher des Deutschen spürt sofort, dass das Kurzwort *als* im Satz

(2) ***Als** ich vor mehr **als** einem Jahrzehnt **als** Kellner arbeitete* […]

jeweils nicht "das gleiche bedeutet", auch wenn er dies nicht unbedingt mit Hilfe von metasprachlichen Termini wie den folgenden explizieren könnte: subordinierende Konjunktion

(Subjunktion), Vergleichspartikel beim Komparativ, Satztteilkonjunktion (Anschlusspartikel bei zugeordneten Satzgliedern).

*DWDS*, *Cosmas tagged* und *Tiger* können die unterschiedlichen Funktionen im Prinzip unterscheiden. Da aber der Annotierung ein ungenügend differenzierter Wortklassenbegriff zugrundliegt (die *POS*-Markierung [= *parts of speech*] beruht weitgehend auf der traditionellen Zehn-Wortarten-Lehre), kann insbesondere die dritte Verwendungsweise nicht als solche erkannt werden, und auch bei der Unterscheidung der beiden erstgenannten sind die entsprechenden Belegzusammenstellungen gespickt mit Fehlanzeigen.

Auch wenn das Grosskorpus *Cosmas* nicht annotiert ist, kann es trotzdem auf vielfältige Weise mittels Anfragezuspitzung für syntaktische Fragestellungen eingesetzt werden. So kann der Benutzer auf Grund seines grammatischen Vorwissens zumindest eine Teilmenge der Belege mit subordinierender Konjunktion *als* herausdestillieren, z.B. durch einen kombinierten Suchbefehl "Komma + *als*" oder durch *Als* mit Grossschreibung (Nebensätze in Voranstellung). Bei *Cosmas tagged* dagegen kann man direkt nach *als* als Subjunktion (10'202 Treffer, bei zahlreichen Fehlanzeigen) oder nach der Kombination "Adjektiv im Komparativ + *als*" (1672 Belege) suchen. Dabei werden nebenbei gesagt auch auf einen Schlag die wichtigsten dabei auftretenden graduierenden Partikeln sichtbar, wie z.B. in *Er ist viel/weit/wesentlich/deutlich/erheblich* usw. *älter als sie*. Beizufügen bleibt, dass gegenwärtig noch kein Suchprogramm imstande ist, semantische Disambiguierungen vorzunehmen, also etwa zwischen der temporalen und der vergleichenden Bedeutung der subordinierenden Konjunktion *als* zu unterscheiden.

Zweites Beispiel: Funktionale Trennung von Flexionsmorphemen

Als Ausgangspunkt soll der folgende Beleg aus dem *Cosmas*-Korpus dienen:

(3) ***Schöner** Orgasmus ist noch **schöner** als ein **schöner** Flug*. (Mannheimer Morgen, 01.08.1995)

Nur *Cosmas tagged* und *Tiger* sind imstande, in einem Beispiel wie (3) zwischen dem flektierten Adjektiv *schön-er* und der unflektierten Adjektivform im Komparativ *schön-er* zu unterscheiden. So findet *Cosmas tagged* etwa den unflektierten Komparativ *schöner* in 76 Belegen und das flektierte Adjektiv im Positiv *schöner* (in verschiedenen Genus-Numerus-Kasus-Kombinationen) in 219 Belegen.

Im – nicht getaggten – *Cosmas*-Grosskorpus kann man allerdings auch hier durch Anfragezuspitzung (zumindest für Teilmengen) zu recht verlässlichen Resultaten gelangen. Die Suchanfrage nach "*schöner* + (direkt gefolgt von) *als*" beispielsweise ergibt 1236 Treffer, von denen praktisch alle die Komparativform enthalten. Umgekehrt werden durch den Suchbefehl "*schöner* NICHT *als*" (alle Belege für *schöner* ohne die Partikel *als* im Kotext) 6127 Belege greifbar, bei denen das Adjektiv überwiegend nicht im Komparativ steht. Aber auch hier ist zur einwandfreien Trennung der beiden homonymen Funktionswörter Handarbeit angesagt.

ISSN 1615-3014

Drittes Beispiel: Funktionale Trennung von Nominalgruppen

Als Ausgangspunkt soll Beispiel (4) dienen, bei dem die gleiche Nominalgruppe einmal als Satzglied und einmal als Attribut verwendet wird.

(4) *Die Alpen bedürfen **des Schutzes** vs. im Interesse **des Schutzes** der Alpen*

Von allen hier vorgestellten Analyseprogrammen ist nur *Tiger* imstande, zwischen der Verwendung von *des Schutzes* als Genitivobjekt oder als Genitivattribut zu unterscheiden. Derartige Disambiguierungen setzen eine syntaktische Analyse aller Sätze des Korpus (*parsing*) und das Vorhandensein von entsprechenden Kategorien in der Abfragesprache voraus (z.B. bei *Tiger* die beiden Marker "AG" vs. "OG"). Allerdings bietet das *Tiger*-Korpus für das – im heutigen Deutsch marginale – Genitivobjekt bloss 162 Treffer, für das Genitivattribut dagegen erwartungsgemäss eine viel grössere Zahl (13'957 Belege). Zu den letzteren sind noch die 1011 mit einer speziellen Suchfunktion abrufbaren vorangestellten Genitivattribute (vom Typ *Schröders Ehefrau*) hinzuzuzählen. Es liegt auf der Hand, dass derartige Suchabfragen für die Satz- und Textsyntax des Deutschen von grösster Relevanz sind.

Auch wenn das Grosskorpus *Cosmas* im engern Sinn keine syntaktischen Suchabfragen erlaubt, so können doch – ausgehend von einem linguistischen Vorwissen – interessante Materialsammlungen erstellt werden (bei deren Bearbeitung dann aber wiederum viel Handarbeit anfällt). Zunächst stellt man mit Hilfe von Grammatiken und anderen Hilfsmitteln (z.B. *Tiger*) eine Liste aller Verben, die den Genitiv regieren, zusammen. Sodann unterwirft man, jeweils ausgehend von einem konkreten Verb wie z.B. *bedürfen*, alle gefundenen Belege einer Kookkurrenzanalyse. Dadurch kann man nicht bloss feststellen, wie frequent Satzbaupläne mit Genitivobjekt im heutigen Deutsch sind, sondern darüberhinaus auch herausfinden, mit welchen Nomen die Position des Genitivobjekts vor allem besetzt ist. Bei *bedürfen* handelt es sich überwiegend um *Zustimmung, Genehmigung, Pflege, Korrektur*, usw. Derartige Angaben sind etwa für die Lexikographie von unmittelbar praktischem Interesse. Insgesamt aber muss gesagt werden, dass *Cosmas* sich hier nur zur punktuellen Überprüfung von Hypothesen, jedoch nicht zur exhaustiven Erfassung von grammatischen Phänomenen eignet.

Viertes Beispiel: Zuordnung von Satzbauplänen zu Verben

Eine konkrete Ausgangsproblematik könnte hier etwa so lauten: Kann das Suchprogramm aus einem Text alle Verben mit einem bestimmten Satzbauplan extrahieren, wenn möglich in einer nach Frequenzen geordneten Liste darstellen? Beispielsweise alle dreiwertigen Verben mit dem – konkret realisierten – Satzbauplan *Subjekt + Akkusativobjekt + Dativobjekt*? Es ist klar, dass auch hier nur noch *Tiger* mithalten kann. Unter den sog. *Three Place Verbs* eruiert *Tiger* alle einschlägigen Belege, darunter etwa die Verben *(jemandem etwas) zeigen/anbieten/liefern/vererben/ verweigern/bescheinigen*.

Aber auch das Grosskorpus *Cosmas* kann hier, wiederum ausgehend von konkreten Verben, interessante Materialsammlungen erstellen, die dann aber teilweise "von Hand" weiterverarbeitet werden müssen. Ausgehend vom dreiwertigen Verb *jemandem etwas zuschanzen*

werden für die Akkusativ-Position u.a. folgende Nomina errechnet*: (lukrative) Aufträge, Posten, Privilegien, Vorteile,* usw.

Fünftes Beispiel: Funktionale Trennung von Satzarten

Alle Textkorpora (mit Ausnahme von *Leipzig*) erkennen die gängigen Satzzeichen wie Punkt, Fragezeichen, Ausrufezeichen und Komma. Sie können also alle Punktsätze (im Sinne von "Sätze mit einem Punkt am Ende und Grossschreibung des folgenden Substantivs"), alle Fragezeichensätze oder alle Ausrufezeichensätze erkennen. Die so gefundenen Teilmengen sind natürlich nicht mit den Deklarativsätzen, Interrogativsätzen, Imperativsätzen usw. gleichzusetzen. Bei *DWDS*, *Cosmas* und *Tiger* können die Satzschlusszeichen auch in kombinierte Suchanfragen eingebaut werden. Darüberhinaus kann *Tiger* (dank *tagging* und *parsing*) die Verbstellungstypen des Deutschen (Verb-Erst-, Verb-Zweit- und Verb-Letzt-Stellung) unterscheiden. Damit können etwa alle Verb-Erstellungs-Belege aus diesem Korpus extrahiert werden (insgesamt handelt es sich allerdings bloss um 433 Treffer). Im Gegensatz zu *Cosmas* muss man hier für einschlägige Suchen also nicht von einem bestimmten Verb ausgehen. Aber auch bei *Tiger* setzt dann spätestens hier die Handarbeit ein, etwa bei einem (fiktiven) Beispiel wie dem folgenden:

(5) (a) *War das eine GUTE Idee?*

(b) *War DAS eine gute Idee!*

(c) *War das eine gute Idee, dann* […].

Eine Trennung zwischen den Verb-Erststellungs-Belegen in der Funktion (a) eines Interrogativsatzes, (b) eines Exklamativsatzes oder (c) eines uneingeleiteten Konditionalsatzes impliziert Kenntnisse der Satzintonation und liegt damit wohl noch für lange Zeit ausserhalb des "Verstehenshorizontes" von automatischen Suchprogrammen. Da aber die Satzzeichen (?/!/,) die Satzintonation zumindest teilweise andeuten, kann man auch hier durchaus zu approximativen Zuordnungen kommen. Allerdings steht etwa bei Exklamativsätzen als Satzschlusszeichen ebenso häufig ein Punkt wie ein Ausrufezeichen, und umgekehrt findet sich das Ausrufezeichen auch in zahlreichen anderen Verwendungen.

Das Grosskorpus *Cosmas* erlaubt auch hier durch die gezielte Verfeinerung der Anfragen (auf Grund von vorgängig per Hand erhobenen Daten) die rasche Erstellung von interessanten Materialsammlungen, die dann wiederum von Hand weiter bearbeitet werden müssen. Wie sich durch die Nachbearbeitung von digitalen Belegsammlungen im Bereich der Satzarten wichtige Erkenntnisse gewinnen lassen, ist Gegenstand des Beitrags von A. Näf in diesem Band. Anhand der dort besprochenen Beispiele soll aufgezeigt werden, wie man mit Hilfe einer Gross-Datenbank wie *Cosmas*, die weder morphologisch noch syntaktisch annotiert ist, durch gezielte Anfragezuspitzung trotzdem zu neuen Einsichten auf dem Gebiet der Syntax der deutschen Gegenwartssprache gelangen kann.

ISSN 1615-3014

## 6 Schluss

Die Wahl einer bestimmten Datenbank als empirischer Grundlage hängt entscheidend vom Ziel einer Anfrage ab. Mit *Leipzig* kommt man bei lexikologischen Fragestellungen auf schnellem Weg zu ersten Ergebnissen. Diese stellen jedoch kaum mehr als erste Richtwerte dar. Der Hauptvorteil dieses Suchprogramms, nämlich die Einfachheit der Anwendung, schränkt eben auch die Einsatzmöglichkeiten dieses Werkzeugs stark ein. Für lexikologische Untersuchungen mit höheren sprachwissenschaftlichen Ansprüchen ist zweifellos *Cosmas* die "beste Adresse", weil es – im Gegensatz zu *Leipzig* – ausgefeilte Analysemethoden (insbesondere Lemmatisierung und Kookkurrenzanalyse mit vielen einstellbaren Parametern) bereitstellt. Zum einen ist das Cosmas-Korpus auch für die Untersuchung von bloss selten auftretenden Wörtern und grammatischen Phänomenen hinreichend umfangreich, zum andern lässt sich umgekehrt bei Suchergebnissen mit grossen Massen von Treffern das Belegmaterial schnell und gezielt einschränken. Ohne Zweifel gehört das Analysewerkzeug *Kookkurrenzanalyse* von *Cosmas* zu den elaboriertesten seiner Art. Ein unschätzbarer Vorteil von *Cosmas* gegenüber allen anderen Korpussuchmaschinen ist zudem die Möglichkeit, dass Zwischenergebnisse weiterverarbeitet und als virtuelle Korpora für weitere Untersuchungen abgespeichert werden können. Bei keinem anderen Suchprogramm kann das nach wie vor intelligenteste und leistungsstärkste Analysewerkzeug überhaupt – nämlich der linguistische Sachverstand des Forschers – so nutzbringend eingebracht werden wie bei *Cosmas*. Dessen Teilkorpus *Cosmas tagged* bietet zudem – ähnlich wie *DWDS* – bestimmte Wortklassen-Suchoptionen an. Dieses Teilkorpus verbindet das grosse Spektrum der Analysemöglichkeiten von *Cosmas* mit dem Vorteil einer morphologischen Annotation und ist von daher bei zahlreichen grammatischen Fragestellungen die richtige Wahl. Die Vorteile von *DWDS* kommen bei Fragestellungen mit diachronischem (20. Jahrhundert) oder domänenspezifischem Untersuchungsziel am besten zur Geltung. Sein grösster Nachteil ist indes, dass man Zwischenergebnisse nicht wie bei *Cosmas* noch weiter bearbeiten kann. Für die Lösung von im engeren Sinne syntaktischen Fragen ist allerdings nur das – gegenwärtig allerdings erst als Demonstrationskorpus verfügbare – *Tiger*-Korpus wirklich geeignet.

Unabhängig davon, ob man die Korpuslinguistik als eigene wissenschaftliche Disziplin oder bloss als linguistische Hilfswissenschaft ansieht, eines scheint klar: Auch wenn die Entwicklung der Abfragemöglichkeiten von digitalen Grosskorpora noch keineswegs abgeschlossen ist, so zeichnet sich doch jetzt schon ab, dass dank der unterdessen erreichten extremen Verkürzung der für die Belegsuche nötigen Zeit ein völlig neuer Blick auf die Sprache möglich geworden ist. Und zum ersten Mal in der Geschichte der germanistischen Sprachwissenschaft kann man nun die berechtigte Hoffnung hegen, dass die künftigen Kodifierungen der deutschen Sprache in Form von Wörterbüchern und Grammatiken die Sprachwirklichkeit vollständig und unverzerrt abbilden werden.

ISSN 1615-3014

**Literaturangaben**

Ammon, Ulrich et al. (eds.) (2004): *Variantenwörterbuch des Deutschen. Die Standardsprache in Österreich, der Schweiz und Deutschland sowie in Liechtenstein, Luxemburg, Ostbelgien und Südtirol*. Berlin.

Belica, Cyril/Steyer, Kathrin (to come): "Korpusanalytische Zugänge zu sprachlichem Usus". Vorabdruck:
www.ids-mannheim.de/kl/projekte/uwv/CBKSpraha.ver20050426.mit.summ.pdf.

Bickel, Hans (2000): "Das Internet als Quelle für die Variationslinguistik". In: Häcki Buhofer, Annelies (ed.): *Vom Umgang mit sprachlicher Variation. Soziolinguistik, Dialektologie, Methoden und Wissenschaftsgeschichte. Festschrift für Heinrich Löffler zum 60. Geburtstag*. Basel: 111-124. (= *Basler Studien zur deutschen Sprache und Literatur* 80).

Habert, Benoît et al. (1997): *Les linguistiques de corpus*. Paris.

Labov, William (1972): *Sociolinguistics Patterns*. Philadelphia.

Lenz, Susanne (2000): *Korpuslinguistik*. Heidelberg. (= *Studienbibliographien Sprachwissenschaft* 32).

Steyer, Kathrin (2004): "Kookkurrenz. Korpusmethodik, linguistisches Modell, lexikografische Perspektiven". In: Steyer, Kathrin (ed.): *Wortverbindungen – mehr oder weniger fest*. Berlin: 87-116. (= *Institut für deutsche Sprache. Jahrbuch* 2003).

ISSN 1615-3014