Web as Corpus: Kooperation mit der Universität Bologna

von Cyril Belica, Holger Keibel, Marc Kupietz und Rainer Perkuhn

Aufgrund der Mannigfaltigkeit und Komplexität sprachlicher Phänomene bedürfen sowohl die Beurteilung existierender sprachwissenschaftlicher Modelle als auch das Aufstellen und Systematisieren neuer Hypothesen über den Sprachgebrauch in der Regel extrem großer Datenmengen. Daher sind Sprachwissenschaftler bemüht, insbesondere auch das World Wide Web mit seinem schier unerschöpflichen Vorrat an sprachlichen Äußerungen als Datenquelle zu erschließen. Die unreflektierte direkte Befragung des Web für sprachwissenschaftliche Zwecke ist jedoch aus wissenschaftsmethodischen Gründen äußerst fragwürdig. Zum Einen verschließt sich das Web gerade wegen der überwältigenden Menge des empirischen Sprachmaterials einer verlässlichen intellektuellen Auswertung über das Niveau von hermeneutischen Deutungen hinaus, zum anderen entzieht es sich auch dem Einsatz von bewährten quantitativen Analysemethoden, weil die distributionellen Eigenschaften der jeweiligen Grundgesamtheit[1] prinzipiell unzugänglich sind.

Die Erschließung des World Wide Web als Datenquelle für die sprachwissenschaftliche Forschung ist also eine keineswegs triviale Aufgabe. Um die grundlegenden und sprachübergreifenden Herausforderungen gemeinsam effektiver überwinden zu können, haben sich einige der Vorreiter auf dem Gebiet Anfang 2005 zu WaCky, der „Web-as-Corpus kool ynitiative", zusammengeschlossen. Im Rahmen dieser informellen Initiative wurden seitdem zahlreiche Methoden und Werkzeuge entwickelt und bereits mit dem Aufbau größerer Web-Corpora begonnen. Auf der EACL[2]-Tagung im April 2006, auf der auch die konstituierende Sitzung der neuen Special Interest Group „Web-as-Corpus" (SIGWAC) der ACL[3] stattfand, stellten Marco Baroni (Universität Bologna) und Adam Kilgarriff (University of Sussex) das erste deutschsprachige Web-as-Corpus deWaC, mit der beachtlichen Größe von 1,7 Milliarden Tokens vor (Baroni & Kilgarriff 2006). Noch auf der Tagung wurde eine Kooperation zum Vergleich von deWaC mit dem Deutschen Referenzkorpus des IDS (DeReKo) vereinbart.

Methodischer Hintergrund

Für Sprachkorpora jeder Art gilt, dass die aus ihnen gewonnenen Erkenntnisse nur dann prinzipiell falsifizierbar sind und auf die jeweilige Grundgesamtheit extrapoliert werden können, wenn die Korpora hinsichtlich der jeweils primären, d. h. sprachwissenschaftlichen Fragestellung als ausgewogen bzw. repräsentativ für diese Grundgesamtheit angenommen werden können. Während eine solche Ausgewogenheit beim Aufbau von traditionellen Textkorpora durch die gezielte Auswahl einer geeigneten Mischung von Texten zu erreichen ist, muss die Zusammensetzung eines durch automatisches Web-Crawling erstellten Textarchivs wie deWaC a posteriori bestimmt werden. Eine solche Bestimmung der Zusammensetzung kann entlang ver-

schiedener extratextueller Dimensionen (z. B. Genre, Thema, Zeit, ...) erfolgen – dies allein reicht jedoch meist nicht aus. Vielmehr müssen die Ergebnisse dieser Bestimmung wiederum in die Crawling[4]-Mechanismen einfließen, um so ggf. in mehreren aufeinander folgenden Sampling[5]- und Bestimmungszyklen die Zusammensetzung der Stichprobe der Zielvorstellung anzugleichen. Weil aber in Unkenntnis der relevanten Grundgesamtheit auch diese Zielvorstellungen a priori nur vage sind, liegt es nahe, die Eigenschaften von DEWAC in Relation zu einem hinsichtlich der oben genannten Dimensionen sehr gut erschlossenen Korpus wie DEREKO zu ermitteln.

Zunächst fungiert DEREKO bei diesen Vergleichen als fester Referenzpunkt: Weicht DEWAC hinsichtlich bestimmter linguistischer Merkmale systematisch von DEREKO ab, so ist analytisch zu klären, inwiefern diese Abweichung eine unerwünschte Eigenschaft von DEWAC darstellt. In solchen Fällen liegt es nahe, die Methodik, mit der das DEWAC-Korpus aus dem Web gewonnen wurde, entsprechend zu verfeinern oder zu erweitern. Umgekehrt können Abweichungen aber auch auf fehlerhafte Texte, Lücken in der Korpuszusammenstellung und andere Defizite in DEREKO hinweisen. Ungeachtet der Schwierigkeit, für eine vorliegende Abweichung zu entscheiden, ob ihr ein Defizit in einem der beiden Korpora zugrunde liegt und wie es ggf. behoben werden kann, ergibt sich als übergreifende Forschungsstrategie, dass die Korpora in mehreren Zyklen verglichen und in angemessener Weise modifiziert werden und sich dadurch in ihrer formalen Qualität und Zusammensetzung iterativ – d. h. schrittweise gegenseitig, auf dem jeweiligen Zwischenstand aufbauend – verbessern. Dabei sollen sich die ersten Vergleichszyklen v. a. auf elementare linguistische Beschreibungsebenen (z. B. Häufigkeitsverteilung von Buchstaben oder Wörtern) konzentrieren, weil Modifikationen eines Korpus auf diesen Ebenen seine Eigenschaften auf übergeordneten Ebenen (z. B. Verteilung der Textthemen oder Textgenres) erheblich beeinflussen können. Frühe Vergleiche auf solchen höheren Ebenen können zwar bereits interessante Unterschiede zwischen beiden Korpora hervorbringen, sind jedoch als sehr vorläufig zu betrachten.

Vergleich DEWAC mit DEREKO

Aufgrund der urheberrechtlichen Beschränkungen wurde im Rahmen der Kooperation bislang auf Vergleiche auf Textebene verzichtet. Stattdessen wurden zu beiden Korpora Tokenlisten erzeugt, auf deren Grundlage sämtliche im Folgenden beschriebenen Vergleiche durchgeführt wurden (ähnliche Vergleiche über Tokenlisten finden sich u. a. bei Sharoff 2006). Unter ‚Token' ist in diesem Sinne eine in sich geschlossene Zeichenkette (ein ‚Wort') zu verstehen. Die Listen enthalten jeweils die Angaben, welche verschiedenen Tokens in den beiden Korpora mit welcher absoluten Häufigkeit vorkommen. Die Tokenlisten wurden jeweils dem anderen Kooperationspartner zur Verfügung gestellt.

Grundsätzliche Vergleichbarkeit

Damit zwei Korpora grundsätzlich vergleichbar sein können, müssen verschiedene Annahmen gelten. Zunächst müssen sie im gleichen Zeichensatz kodiert sein, und idealerweise ist ihr Umfang annähernd gleich groß. Diese beiden Voraussetzungen waren erfüllt. Für die Vergleichbarkeit der Tokenlisten kommt hinzu, dass die Listen nach den gleichen Prinzipien bzw. mit denselben Verfahren aus dem jeweiligen Korpus abgeleitet werden müssen. Um dies gewährleisten zu können, stellte die Universität Bologna dem IDS die Methoden zur Verfügung, mit der die DEWAC-Tokenliste erzeugt worden war. Diese Verfahren wurden zur Generierung der IDS-Tokenliste auch auf DEREKO angewandt. Eine weitere Annahme steht in einem sehr engen Bezug zum eigentlichen Gegenstand der Untersuchung. Sprachliche Phänomene manifestieren sich in Texten mit Eigenschaften auf verschiedenen Dimensionen. Damit überhaupt vom Korpus auf die Phänomene in der jeweiligen Grundgesamtheit geschlossen werden kann, muss die Zusammensetzung des Korpus hinsichtlich der für die jeweilige Fragestellung relevanten Dimensionen (z. B. Textgenre, Textthema) die Zusammensetzung der Grundgesamtheit widerspiegeln. Während die Zusammensetzung von DEREKO entlang einiger Dimensionen bekannt ist, gilt dies für DEWAC nicht. Genauso wenig ist bekannt, wie sich die beiden Grundgesamtheiten, der allgemeingeläufige Gebrauch des Deutschen und der Gebrauch des Deutschen im World Wide Web, zueinander verhalten. Bei der Interpretation der Auffälligkeiten, die bei den Vergleichen hervortreten, sind deshalb verschiedene mögliche Ursachen zu beachten. Die Abweichungen können durch das Vorgehen bei der Korpuserstellung oder durch die unterschiedlichen Zusammensetzungen der Grundgesamtheiten bedingt sein.

Im Rahmen dieser Kooperation wurden DEWAC und DEREKO u. a. hinsichtlich ihrer Häufigkeitsverteilungen auf drei verschiedenen Ebenen verglichen: auf der Zeichenebene, auf der Tokenebene und tentativ auf der Themenebene. Diese Vergleiche werden nachfolgend in Auszügen vorgestellt.

Vergleich auf Zeichenebene

Wenn die beiden Korpora, so wie skizziert, über Tokenlisten vergleichbar sein sollen, so müssten sie auf Zeichenebene ungefähr dieselbe Häufigkeitsverteilung aufweisen. Eine einheitliche Tokenisierung[6] war zwar gewährleistet, es könnten sich aber Auffälligkeiten aufgrund der Unterschiede in den Grundgesamtheiten oder bei den Techniken und Konventionen bei der Textaufbereitung niedergeschlagen haben. Gerade unterschiedliche Aufbereitungskonventionen können durchaus weitreichende Folgen haben, wenn z. B. die Kodierung von token-begrenzenden Zeichen (wie Leerzeichen, Satzzeichen, Klammerzeichen usw.) unterschiedlich gehandhabt wird. Daran wird deutlich, dass es keinen Konsens darüber gibt, wie das Konzept eines ‚(Roh-)Textes' als Input für ein Korpus verstanden und auch ausgelegt werden kann, weder dahingehend, was einen Text als solchen auszeichnet (und ihn von anderen abgrenzt), noch darüber, wie die Struktur der internen Bestandteile zu fassen ist. Ein erstes – und fast das wichtigste – Anliegen der Kooperation war deshalb, diese Konzeption und die Aufbereitungsverfahren speziell vor dem Hintergrund dieses Vergleichs zu durchleuchten und in Zukunft soweit wie möglich aufeinander abzustimmen, um sie als vermeintliche Störquellen auszuschließen. Denn nur dann wäre es möglich, Aussagen über die Grundgesamtheiten zu formulieren.

Indizien / Diskussion

Auf der Zeichenebene wurden lediglich die Zeichen näher untersucht, deren Vorkommenshäufigkeiten in den beiden Korpora stark voneinander abwichen. Die größten Unterschiede betrafen Satzendezeichen und Zeichen wie ‚<', die in den struktur-auszeichnenden Elementen der Korpustexte verwendet werden und mutmaßlich durch eine ungenügende Abstimmung zwischen Rohdaten, Aufbereitungskonventionen und Tokenisierungsmethoden in die Wortlisten gelangt sind. Starke Häufigkeitsabweichungen wurden außerdem für einige nicht-alphanumerische Zeichen festgestellt, die in DEWAC uneinheitlich kodiert sind, was wahrscheinlich eine Folge fehlender oder falscher Zeichensatz-Angaben in den zugrunde liegenden Webseitenquelldaten ist. Diese Auffälligkeiten lassen sich also vermutlich eher auf technische als auf inhaltliche Gründe zurückführen; um dies zu überprüfen, sollten die Experimente aber nach einer Überarbeitung der Verfahren wiederholt werden.

Da signifikante Unterschiede festgestellt wurden, galt für alle weiteren Untersuchungen, dass sie unter Vorbehalt durchgeführt wurden. Solange die Ursache nicht geklärt ist, muss das vorrangige Ziel sein, die Aufbereitungskonventionen so lange zu vereinheitlichen, wie sie potenziellen Einfluss haben. Trotzdem erschien es sinnvoll, die Untersuchungen fortzuführen, zum einen, um weitere Erfahrungen mit der Methodologie zu sammeln, zum anderen, um – wenn auch unter Vorbehalt – weitere Erkenntnisse aus dem Vergleich ableiten zu können.

Vergleich auf Tokenebene

Auf der Tokenebene war die Häufigkeitsverteilung von Wörtern und wortähnlichen Zeichenketten Gegenstand der Untersuchung. Ähnliche Korpora sollten auch in dieser Hinsicht ähnliche Verteilungen aufweisen. Dabei war selbstverständlich nicht zu erwarten, dass sich Häufigkeitsränge oder relative Frequenzen exakt entsprechen würden. Andererseits wären starke Abweichungen genauso überraschend. Von besonderem Interesse wären systematische Abweichungen bezüglich textsorten-konstituierender Eigenschaften. Inwieweit eine Abweichung des erwartbaren Wertes noch innerhalb einer Spanne zufälliger Einflüsse liegen kann, lässt sich mithilfe statistischer Maße bewerten. In diesem Experiment wurde die ‚log-likelihood ratio'[7] verwendet (vgl. Dunning 1993) und für jede Tokenhäufigkeit in DEWAC bezogen auf dessen Häufigkeit in DEREKO berechnet. Die Tokens, die am stärksten über- bzw. unterrepräsentiert waren, wurden exemplarisch von Hand ausgewertet. Diese deutlichsten Abweichungen sollten einen ersten Eindruck über die Zusammensetzung der beiden Korpora liefern.

Indizien / Diskussion

Eine erste flüchtige Auswertung zeigt, dass in DEWAC Personal- und andere Pronomina in der ersten und zweiten Person (*ich, du, Du, mir, mich, wir, Ich, uns, dir, euch, dich, meine, Dir, deine*) und entsprechende Modal-/Auxiliarverben im Präsens (*kann, hast, muss, können*) stark vertreten sind. Weiterhin spezifisch für das DEWAC sind Bezeichnungen aus dem juristischen und religiösen Umfeld (*Klägerin, Beklagten, Kläger, Gott, Jesus*), Abkürzungen (*BGB, vgl., BStBl, EStG*), außerdem ist eine leichte Tendenz zu umgangssprachlichen Ausdrücken zu verzeichnen. Typisch für DEREKO sind hingegen Personalpronomina in der dritten Person (*er*), aber nicht so durchgängig wie komplementär bei DEWAC, sowie entsprechende Verbformen, häufig im Konjunktiv bzw. in Vergangenheitsformen (*sei, sagte, seien, will, werde, worden, erklärte*). Desweiteren fallen auf als typisch für DEREKO: Zeitangaben (*gestern, Sonntag, Samstag, Montag* [usw.], *Jahren*), Währungsangaben (*Mark, Schilling, Franken, Dollar*), Zahlen und Größenordnungen (*Millionen, Milliarden, zwei, zehn, drei, fünf, vier* [usw.]), Politisches und Ämter (*SPD, CDU, ÖVP, Bürgermeister*) sowie Ortsangaben (*Schweizer, Wiener, Österreich*).

Zusammengefasst sind offenbar Bezeichnungen aus den Bereichen Wirtschaft und Politik charakteristisch für DeReKo, während sich auf Seiten von deWaC die Themen Rechtswesen und Religion abzeichnen. Darüber hinaus lassen sich die Indizien versuchsweise dahingehend interpretieren, dass der Stil der deWaC-Texte eher erzählend ist, den Leser direkt ansprechend, und dass die Texte eine situative Kommunikation wiedergeben, während die DeReKo-Texte eher berichterstattenden Charakter haben. Die Webtexte sind in dieser Hinsicht eher vergleichbar mit mündlicher als mit schriftlicher Kommunikation. Dies bestätigt ein Abgleich der Untersuchung mit einem Korpus gesprochener Sprache (Pfeffer-Korpus[8]): Die für dieses Korpus besonders typischen Wörter sind bis auf eine Ausnahme auch deWaC-typische Wörter. Die Ausnahme stellt das Wort *daß* dar. Dieses Wort wurde bei der Verschriftlichung des Pfeffer-Korpus nach der alten Rechtschreibregelung noch als *daß* transkribiert. Da DeReKo Texte über einen langen Zeitraum abbildet (seit 1964), tritt diese Schreibweise auch dort häufig auf – im Gegensatz zu deWaC, wo die Form *dass* überrepräsentiert ist. Dies kann daran liegen, dass die große Mehrheit der Texte jüngeren Datums ist und die Schreiber sich an die neuen Rechtschreibregeln gehalten haben oder dass sie die Form in vorauseilendem Gehorsam und vor allem deshalb gewählt haben, weil sie für die Eingabe auf dem Computer schon lange der anderen Form vorgezogen wurde.

Vergleich auf Themenebene

In einer allerersten Annäherung sollte auch die thematische Verteilung innerhalb der verschiedenen Korpora untersucht werden. Eine thematische Verteilung müsste sinnvollerweise anhand von Texten analysiert werden. Aufgrund der genannten urheberrechtlichen Einschränkungen wurde stattdessen ein Experiment ausgehend von den Tokenlisten konstruiert. In gewisser Weise sollte sich die thematische Beschaffenheit eines Korpus auch in seiner Tokenliste niederschlagen (erste Beispiele hierfür wurden bereits oben genannt). Abweichungen sind zwar zu erwarten, eine systematische Abweichung ist jedoch ein Indiz für eine unterschiedliche Vorgehensweise bei der Korpuskomposition oder für Unterschiede bei den Grundgesamtheiten. Für den ersten Fall ist das langfristige Ziel die Einsicht in den Zusammenhang von Abweichung und Komposition, um bei Bedarf gegensteuern zu können, für den zweiten Fall die Dokumentation der Unterschiede.

Für dieses Experiment wurde ein Maß entwickelt, das die Themenzugehörigkeit eines Tokens ausdrücken soll. Die meisten DeReKo-Texte sind thematisch klassifiziert, wobei ein Zahlenwert zwischen 0 und 1 für jeden Text gewichtet, wie sicher sich der Klassifikator bei dieser Themenzuweisung war (vgl. Weiß 2005). Abhängig von der Anzahl der Texte, in denen ein Token vorkommt, und den jeweiligen Themen-Gewichten, wurde ermittelt, welchem Thema ein Token am wahrscheinlichsten zuzuordnen ist. Dieses erste Maß wurde gemäß der relativen Häufigkeit standardisiert, für die jeweils hundert typischsten Wörter in DeReKo bzw. deWaC berechnet, über diese hundert Wörter gemittelt und bezogen auf die Themen kumuliert. Die Themen wurden gemäß der Differenz der Maßzahl sortiert. Die Abbildung unten zeigt das obere (‚deWaC-lastige') und das untere (‚DeReKo-lastige') Ende der sortierten Themen.

DIFF	dewac	dereko	topic category
10.32	12.71	2.39	Staat_Gesellschaft:Biographien_Interviews
3.15	4.96	1.81	Kultur:Literatur
2.51	4.66	2.15	Staat_Gesellschaft:Recht
2.47	4.71	2.24	Staat_Gesellschaft:Familie_Geschlecht
1.75	5.63	3.88	Freizeit_Unterhaltung:Reisen
1.69	4.09	2.40	Kultur:Musik
1.46	2.59	1.13	Kultur:Film
0.99	2.49	1.49	Staat_Gesellschaft:Kirche
0.72	1.55	0.82	Technik_Industrie:EDV_Elektronik
0.69	0.75	0.06	Staat_Gesellschaft:Tod
:	:	:	:
:	:	:	:
:	:	:	:
-0.60	0.16	0.76	Wirtschaft_Finanzen:Waehrung
-0.98	3.87	4.85	Freizeit_Unterhaltung:Vereine_Veranstaltungen
-1.04	2.15	3.19	Wirtschaft_Finanzen:Sozialprodukt
-1.17	0.95	2.12	Sport:Vermischtes
-1.84	0.98	2.83	Staat_Gesellschaft:Verbrechen
-2.01	0.82	2.83	Technik_Industrie:Unfaelle
-2.43	7.13	9.56	Politik:Kommunalpolitik
-2.72	1.76	4.48	Wirtschaft_Finanzen:Oeffentliche_Finanzen
-3.42	5.54	8.96	Politik:Ausland
-3.65	3.94	7.59	Sport:Fussball
-4.95	5.75	10.70	Politik:Inland

Abb. 1: Quantitative Annäherung der thematischen Zusammensetzung von deWaC und DeReKo (Ausschnitt)

Indizien / Diskussion

In deWaC ragt das Thema ‚Biographien/Interviews' heraus, ansonsten sind die Themen ‚Literatur/Musik/Film', ‚Recht/Familie/Kirche/Tod', ‚Reisen', und auch noch ‚EDV/Elektronik' gut vertreten. Typisch für DeReKo sind hingegen ‚Politik: Inland/Ausland/Kommunalpolitik', ‚Wirtschaft/Finanzen: Öffentliche Finanzen/Sozialprodukt/Währung', ‚Sport: Fußball/Vermischtes', aber auch ‚Verbrechen', ‚Technik/Industrie: Unfälle'. Dies bestätigt in weiten Zügen die Erkenntnisse aus dem vorangegangenen Vergleich auf Tokenebene: Die Erzählperspektive entspricht den Biographien/Interviews, Juristisches und Religiöses taucht ebenfalls wieder auf. Umgekehrt dokumentieren die Themen ‚Wirtschaft', ‚Politik' und ‚Finanzen' die Berichtslastigkeit von DeReKo. Ein Kontrollexperiment mit künstlich generierten Texten, die jeweils nur aus Wiederholungen der typischen Tokens bestanden, bestätigte diese Ergebnisse. Das Thema ‚EDV' wurde bei diesem Versuch aber noch ausgeprägter als typisch für deWaC eingestuft.

Auswertung

Die ersten Auswertungen im Rahmen der Kooperation liefern bereits wertvolle Erkenntnisse, wohlgemerkt allerdings nur über den aktuellen Zustand der beiden Korpora. Der hier skizzierte kurze Überblick soll nicht davon ablenken, dass dies nicht das eigentliche Anliegen der Kooperation war. Viel wesentlicher war, Erfahrungen über Metriken und Methoden zu sammeln, mit denen grundsätzlich die Zusammensetzung von Korpora bestimmt werden kann, und wie diese Erkenntnisse langfristig in Methoden oder Parameter zum kontrollierten Korpusaufbau umgesetzt werden können.

Mit einfachen Mitteln lässt sich kein Korpus unkontrolliert aus dem World Wide Web extrahieren, mit dessen Hilfe sich Aussagen über den allgemeinen Sprachgebrauch formulieren ließen. Dies mag daran liegen, dass sich im Web ein eigener Sprachgebrauch manifestiert. Oder es mag daran liegen, dass zurzeit noch die Mittel fehlen, einen ‚repräsentativen' Ausschnitt aus dem Web zu erheben. Doch unabhängig von den tatsächlichen Gründen wäre der nächste unabdingbare Schritt, die Eigenschaften der Texte in den Dimensionen genauer zu erforschen, die zur Ausprägung sprachlicher Phänomene beitragen. Diese Vorstudie war erst ein kleiner Schritt in diese Richtung.

Anmerkungen

[1] „In der empirischen Forschung bezeichnet die Grundgesamtheit (auch *Population*) die Menge aller potentiellen Untersuchungsobjekte für eine bestimmte Fragestellung. Aus pragmatischen Erwägungen wird normalerweise nicht die Grundgesamtheit, sondern eine Stichprobe untersucht, die für die Grundgesamtheit repräsentativ ist." (wikipedia.de, Januar 2007)

[2] European Chapter of the Association for Computational Linguistics.

[3] Association for Computational Linguistics.

[4] Web-Crawler (oder auch Web-Spider oder Web-Robots) sind Verfahren des automatischen Durchwanderns des World Wide Web. Ausgehend von einer Startmenge von Webseiten wird wiederholt versucht, Verknüpfungen zu weiteren Seiten zu identifizieren, die wiederum weiter verfolgt werden. Von den dadurch erfassten Seiten werden z. B. für Suchmaschinen oder für Webkorpora Kopien gesammelt, d. h. es wird auf die Seiten nur lesend und nicht manipulierend zugegriffen.

[5] Vorgang bzw. Ergebnis beim Ziehen einer Stichprobe aus einer Grundgesamtheit.

[6] Bestimmung der Tokens eines Textes, d. h. der wortähnlichen Zeichenketten, die für weitere Betrachtungen des Textes relevant sind. Dabei ist es allerdings nicht unproblematisch zu entscheiden, welche Zeichen zu einem Token dazugehören (ein Punkt am Satzende vs. nach einer Abkürzung), welche Zeichen verschiedene Tokens trennen (Leerzeichen vs. Bindestrich), welche Zeichen evtl. sogar verschiedene Bestandteile eines Tokens verbinden (Worttrennung am Zeilenende, elliptische Formulierungen).

[7] Statistisches Maß zur Bewertung von Hypothesen auf der Grundlage empirischer Daten.

[8] <http://dsav-wiss.ids-mannheim.de/DSAv/KORPORA/PF/PF_DOKU.HTM>

Literatur

Baroni, M./Kilgarriff, A. (2006): Large linguistically-processed web corpora for multiple languages. EACL 2006 Conference Companion, S. 87-90.

Dunning, Ted (1993): Accurate Methods for the Statistics of Surprise and Coincidence. In: Computational Linguistics 19 (1), S. 61-74.

Sharoff, S. (2006): Creating general-purpose corpora using automated search engine queries. In: Baroni, M./ Bernardini, S. (Hgg.), WaCky! Working papers on the Web as Corpus. <http://wackybook.sslmit.unibo.it>. Bologna: Gedit, S. 63-98.

Weiß, Christian (2005): Die thematische Erschließung von Sprachkorpora. Mannheim: Institut für Deutsche Sprache. (= OPAL – Online publizierte Arbeiten zur Linguistik, 1/2005)

Die Autoren sind wissenschaftliche Mitarbeiter am Institut für Deutsche Sprache in Mannheim.