# DAS INTERNET ALS KORPUS?
# AKTUELLE FRAGEN UND METHODEN DER KORPUSLINGUISTIK

Stefan Diemer, TU Berlin, Germany

Der Artikel beschreibt aktuelle korpuslinguistische Methoden und Anwendungen. Nachfolgend wird die Möglichkeit der Verwendung internetbasierter Daten als Quelle für Korpusforschung wird. Die Grenzen eines solchen internetbasierten Ansatzes werden aufgezeigt. Auf der Basis der aktuellen Entwicklungen im Bereich der Suchmaschinentechnologie erfolgt ein Ausblick auf mögliche zukünftige Anwendungen des Internets als Korpus.

SCHLAGWÖRTER: Korpuslinguistik, Internetbasierte Korpora, Suchtechnologie, Sprachwandel, Lexikologie

## 1 KORPUSLINGUISTIK UND DIE THEORETISCHE SPRACHWISSENSCHAFT

Ein Korpuslinguist trifft einen theoretischen Linguisten. Sie unterhalten sich, und der Korpuslinguist fragt schließlich: „Warum sollte ich glauben, dass das wahr ist, was Sie mir erzählen?“ Der Theoretiker entgegnet: „Warum sollte ich glauben, dass das interessant ist, was Sie mir erzählen?“ (und lehnt sich in seinem Sessel zurück)

Dieser Scherz des Linguisten Charles Fillmore (1992: 35) zeigt gut das Spannungsfeld zwischen deskriptivem und theoretischem Ansatz, zwischen Performanz und Kompetenz (wie Geoffrey Leech 1992 diese beiden Pole bezeichnete). Das Bild: statt zu interpretieren, verlassen sich Korpuslinguisten und -linguistinnen blind auf die Datenmasse. Sie produzieren computerlesbare Textsammlungen und haben anschließend Schwierigkeiten bei der Interpretation.

Tatsächlich ist das Bild der Korpuslinguistik, einer relativ jungen linguistischen Disziplin, noch sehr stark von dieser Vorstellung geprägt. Die ersten Korpora wurden von der etablierten Linguistik anfänglich oft als nutzlos eingestuft, wie W. Nelson Francis und Henry Kucera, die beiden Pioniere des Brown Corpus, feststellen mussten. Noch lange Zeit danach wurde Korpuslinguistik als bloße Hilfswissenschaft und Lieferant von Beweisen für linguistische Theorien gesehen. Noam Chomsky (1995: 170ff.) sprach der Korpuslinguistik nur die Fähigkeit zu, die beiden ersten Stufen seines Adäquanzmodells (Beobachtung und Beschreibung) zu erfüllen. Sie sei aber nicht in der Lage, die höchste Stufe, nämlich die adäquate Erklärung, zu erreichen, gehe also am Kern der Sprache vorbei.

Dieses anfängliche Misstrauen ist erfreulicherweise vor dem Hintergrund moderner korpuslinguistischer Arbeit und der heutigen technischen Möglichkeiten gewichen. Es stimmt, dass viele Korpuslinguisten und -linguistinnen skeptisch gegenüber rein

theoretischen Diskussionen sind, die sich nicht auf Beobachtung gründen. In den letzten dreißig Jahren, und ganz besonders in den letzten zehn, hat sich mit dem Fortschritt der Technik aber die Korpusarbeit wesentlich gewandelt. Korpuslinguisten und Linguistinnen sind heute aktiv in der theoretischen Forschung tätig und ohne Weiteres bereit, an die Datensammlung eine Interpretation der Ergebnisse anzuschließen. Im Folgenden soll diese Entwicklung kurz skizziert und Beispiele für Anwendungsbereiche von Korpusarbeit gegeben werden. Es werden einige Projekte vorgestellt, die die verfügbaren Möglichkeiten innovativ nutzen, um im Abschluss aufzuzeigen, wo die Korpusarbeit an ihre Grenzen stößt, und wie ihre Zukunft aussehen könnte.

## 2 Entwicklung und Anwendungsgebiete

### 2.1 Entwicklung der Korpusforschung

Die moderne Korpuslinguistik entstand erst in den 60er Jahren. Die Sammlung von Sprachdaten ist natürlich weit älter, aber vor dem Aufkommen elektronischer Texte war dieser Prozess enorm zeitraubend und schwierig; außerdem war meist eine statistische Relevanz nicht gegeben. Besonders beeindruckend ist die Arbeit einer Gruppe deutscher Linguisten, die um 1900 Sprachdaten über die Entwicklung englischer Partikeln gesammelt und per Hand quantifiziert haben, meist eine Partikel pro Dissertationsvorhaben, z.B. Grimm und Gasner. Und faszinierend ist natürlich auch die Entstehungsgeschichte des Zitatkorpus für das Old English Dictionary, bei dem über Jahrzehnte hinweg Korrespondenten Textstellen per Hand niederschrieben und an die Redaktion weitergaben.

In den späten 50er Jahren entstand dann erstmals der Bedarf nach quantifizierbaren Sprachdaten. Randolph Quirk und andere entwickelten vor diesem Hintergrund das Survey of English Usage (SEU) mit etwa einer Million Wörtern. In den 60ern kompilierten Henry Kucera und Nelson Francis das eine Million Wörter umfassende Brown-Korpus für amerikanisches Englisch, das für lange Zeit eine Standardquelle für Korpusanalysen wurde.

Abbildung 1: Ein Auszug aus dem Brown-Korpus von 1960

- A. PRESS: Reportage (*44 texts*)
- B. PRESS: Editorial (*27 texts*) (...)
- D. RELIGION (*17 texts*)
- E. SKILL AND HOBBIES (*36 texts*)
- F. POPULAR LORE (*48 texts*)
- G. BELLES-LETTRES - Biography, Memoirs (*75 texts*)

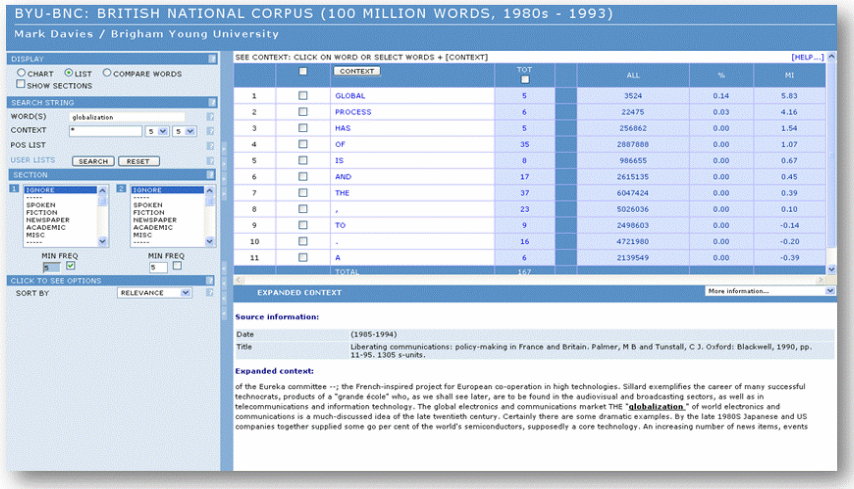
(...)

(Brown Corpus Manual: http://khnt.aksis.uib.no/icame/manuals/brown/)

In den 70ern wurde ihm mit dem London-Oslo-Bergen Korpus (LOB) von Stig Johansson und vielen anderen ein entsprechendes Korpus für britisches Englisch zur Seite gestellt, und mit der Veröffentlichung des London-Lund Korpus durch Jan Svartvik existierte auch ein Korpus für gesprochenes Englisch.

Die 80er und 90er waren gekennzeichnet durch eine rasante technologische Entwicklung. Auf der Basis der existierenden Korpora entstanden eine Vielzahl von Studien zu Einzelphänomenen, z.B. John Sinclairs einflussreiches *Corpus, Concordance, Collocation* (1991) und Michael Stubbs' Untersuchungen zur lexikalischen Semantik. Auch größere Vergleichs- und Folgekorpora wurden entwickelt, wie z.B. FLOB und FROWN an der Uni Freiburg. Für historische Studien entstand das diachronische Helsinki-Korpus. In den 2000ern schließlich prägt die massenhafte Verfügbarkeit von computerlesbaren Texten durch das Wachstum des WWW die Korpusforschung. Grammatische und andere sekundäre Informationen können zunehmend automatisch integriert werden. Die Korpora werden größer und repräsentativer – ein gutes Beispiel ist das British National Corpus mit 100 Millionen Wörtern vieler Textgattungen inklusive gesprochener Sprache.

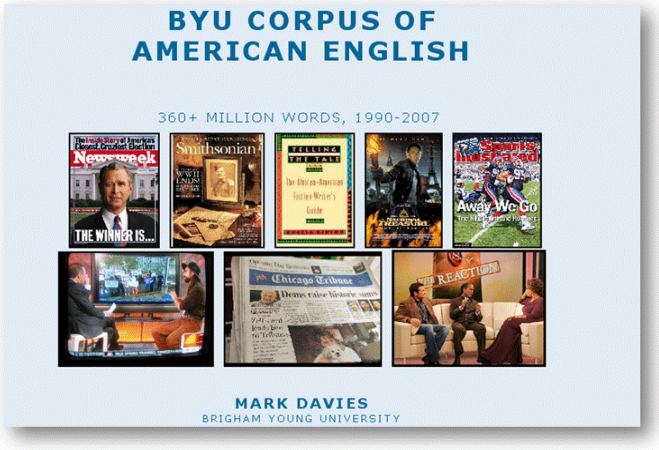
Abbildung 2: Recherche im BNC




(BNC: http://corpus.byu.edu/bnc/x.asp)

Oben abgebildet eines der online verfügbaren Suchportale für das BNC. Die Suchmaske auf der linken Seite erlaubt die Angabe von bis zu zwei Suchbegriffen, die Spezifizierung von Textarten innerhalb des Korpus und die Formatierung des Ergebnisses – im Beispiel der Kontext des Begriffs *globalisation*. Noch größer ist das 360 Millionen Wörter umfassende Corpus of American English, kompiliert von Mark Davies. Es ist wie das BNC online verfügbar und bietet zahlreiche Anwendungsmöglichkeiten in Forschung und Lehre. Allerdings sind noch nicht alle Urheberrechtsfragen geklärt, so dass die Verfügbarkeit unter Umständen wieder eingeschränkt werden könnte.

Abbildung 3: Titel des BYU-CAE




(http://www.americancorpus.org/)

Man sollte meinen, dass mit einer solchen Datenmenge alle Wünsche erfüllt seien. Doch im Winter 2006 gab die Firma Google, die die weltweit beliebteste Internet-Suchmaschine betreibt, ein neues Korpus mit einer Billion Wörtern heraus, das Google Korpus. Leider sind die Daten nur als n-gram-Korpus, also in bis zu 5 Wörtern langen Stücken, verfügbar, was die Nutzbarkeit z.B. im Bereich Syntax einschränkt.

Abbildung 4: Google n-gram

Auszug der 4-gram-Datensätze:

- serve as the incoming 92
- serve as the incubator 99
- serve as the independent 794
- serve as the index 223
- serve as the indication 72
- serve as the indicator 120
- serve as the indicators 45
- serve as the indispensable 11
- serve as the indispensible 40
- serve as the individual 234
- serve as the industrial 52
- serve as the industry 607
- serve as the info 42
- serve as the informal 102
- serve as the information 838
- serve as the informational 41
- serve as the infrastructure 50
- serve as the initial 5331
- serve as the initiating 125

(http://www.ldc.upenn.edu/Catalog/)

Mit einem solchen Korpus ist es möglich, ganz andere Ergebnisse zu erhalten. Ein Hauptproblem von geringen Datenmengen ist die geringere statistische Relevanz. Schlimmstenfalls werden Suchbegriffe nur ein oder zweimal gefunden, und die Aussagekraft solcher Hapaxe (Einmalvorkommen) ist begrenzt. In einer Billion Wörtern

finden sich dagegen sogar Fehler und andere Varianten. Auch die Methodik hat sich geändert. Computer wurden erst Anfang der 90er Jahre wirklich zur Standardausrüstung. In den ersten Jahren der Korpusarbeit mussten, wie sich viele ehemalige studentische und wissenschaftliche Mitarbeiter noch erinnern werden, Bücher per Hand abgetippt oder zerlegt und eingescannt werden – etwa 2 Minuten pro Seite. Anschließend mussten die Fehler (die Erkennungsrate war um 95%) manuell verbessert werden. Dieses Korpus dann zu durchsuchen und weiterzugeben, war ebenfalls schwierig. Geoffrey Leech (1998: xvii) betonte vor zehn Jahren noch, dass in der Korpuslinguistik „eine Menge harter Arbeit geleistet werden muss, bevor die Forschungsresultate eingefahren werden können". Um nur ein Beispiel zu nennen: 1994 wurde im Rahmen eines Forschungsvorhabens ein 3 Millionen Wörter großes Korpus aus mittelenglischen Texten erstellt (Wycliffe-Korpus, 1994-98, siehe auch Diemer 1998). Es dauerte zwei Jahre, diese Texte elektronisch aufzubereiten, und nochmals zwei Jahre für die Analyse mit speziellen Rechercheprogrammen.

Heute, wieder 10 Jahre später, sind Korpora nicht nur als linguistische Datensammlungen, sondern auch über Internetquellen verfügbar, und sie lassen sich mit zunehmender Leichtigkeit und Schnelligkeit durchsuchen. Eine ähnliche Analyse wie damals wäre heute weit weniger aufwändig. Gleichzeitig ist das Problem der Korpuserstellung erheblich reduziert worden. Statt mühsam spezielle Daten (z.B. aus dem Bereich Sportkommentar) zu sammeln, kann nun ein existierendes Korpus durchsucht oder leicht eines erstellt werden. Hier ein Beispiel für die Untersuchung unterschiedlicher Terminologie im amerikanischen Präsidentschaftswahlkampf anhand eines Korpus aus Internetquellen. Im Rahmen eines Korpuslinguistik-Workshops an der Universität des Saarlandes untersuchten Schwarz und Wilke (2008) den Kontext, in dem das Wort *war* in einem Korpus aus politischen Reden der beiden Konkurrenten Hillary Clinton und Barack Obama verwendet wurde. Die Kollokationen lassen sowohl qualitativ wie quantitativ auf einen unterschiedlichen Gebrach schließen, was durch eine Clusteranalyse der zusammenhängenden Ausdrücke mit *war* als Element und durch eine Plot-Analyse, die die Verteilung im Text der Reden untersucht, untermauert wird.

Abbildung 5: Beispiel aus einem webbasierten Korpus: Clinton / Obama


(Schwarz u.a. 2008)

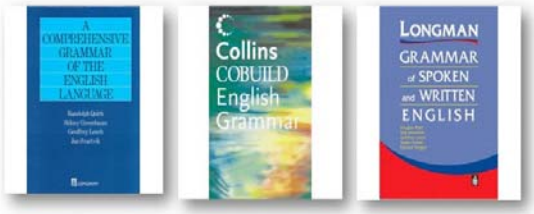
## 2.2 AKTUELLE ANWENDUNGEN

In den folgenden Abschnitten werden in der von Charles F. Meyer (2002) vorgeschlagenen Taxonomie typische Fragestellungen, Methoden und Anwendungsmöglichkeiten der modernen Korpuslinguistik skizziert.

### 2.2.1 GRAMMATISCHE STUDIEN UND REFERENZGRAMMATIKEN

Bei grammatischen Studien und der Kompilierung von Referenzgrammatiken ist der Vorteil eines Referenzkorpus klar: anstelle von konstruierten Sätzen können Beispiele verwendet werden, die dem aktuellen Sprachgebrauch entstammen. Zusätzlich ist es möglich, die Häufigkeit und Akzeptanz dieser Formen zu belegen. Das Hauptproblem bei Studien zur Grammatik war lange der Mangel an Beispielen in den relativ kleinen Korpora. Zwar kann man auch Einzelbeispiele analysieren, aber der Hauptvorteil des Korpus, die quantitative Komponente, entfällt. Das hat sich in den letzten Jahren geändert. Statt, wie Jennifer Coates 1983, die englischen Modalverben mit Hilfe des Brown-Korpus und einer Million Wörtern zu untersuchen, konnte Dieter Mindt 1995 ein 80 Millionen Wörter großes Korpus für seine empirische Grammatik des englischen Verbs verwenden. Wieder zehn Jahre später kann nun leicht ein mehrere hundert Millionen Wörter umfassendes Vergleichskorpus erstellt werden. Seit 2000 sind daher zahlreiche Studien entstanden, die grammatische Konstruktionen mit Hilfe sehr großer Korpora untersuchen. Stellvertretend sollen nur zwei neuere genannt werden: die Analyse des Get-Passivs mit Hilfe des BNC durch Christoph Rühlemann 2007, oder Kristin Davidses Untersuchung der Funktionen von *kind, sort, type* in der Nominalphrase 2008. Mit größeren Korpora können mehr Beispiele und Varianten gefunden und auch quantifiziert werden. Korpusbasierte Referenzgrammatiken sind nichts wirklich Neues: Schon Otto Jespersen benutzte in der ersten Hälfte des letzten Jahrhunderts für seine historische englische Grammatik literarische Quellen – die Kompilation dauerte allerdings auch mehrere Jahrzehnte. Heute setzten viele große Verlage auf Grammatiken, die authentische Korpusbeispiele verwenden.

Abbildung 6: Korpusbasierte Grammatiken




• Quirk / Greenbaum • Collins COBUILD • Longman

(www.amazon.de)

Berühmt ist Randolph Quirks und Sidney Greenbaums Grammatik der englischen Sprache und ihre Nachfolger auf der Basis des ICE-GB Korpus. Collins COBUILD bauen ihre Referenzgrammatiken auf dem von ihnen erstellten Bank of English Korpus mit inzwischen über 500 Millionen Wörtern auf.

Auch Longman benutzt Korpora für seine Grammatiken, so z.B. die Longman Grammar of Spoken and Written English (1999-2004) von Douglas Biber (und Stig Johansson und Geoffrey Leech, um nur die wichtigsten Autoren zu nennen). Ein großer Vorteil für Sprachenlernende sind dabei die authentischen Beispiele, wie hier:

Abbildung 7: Gegenüberstellung von Beispielsätzen zum *if*-Satz

- Nicht korpusbasiert
  - If you do not finish on time, you will fail the exam.
  - If you had a bit of sense, you would be better off now.
  - If you were here, I would be glad.

(Englischunterricht 9. Klasse Gymnasium)

- Korpusbasiert
  - If you want an idea of what is going on, look at Figure 7.
  - And if you had a democratically elected house it would totally alter the balance of the Lords.
  - However, if you were completely rational, you would admit the possibility.

  (drei von fast 50 000 möglichen Beispielen)

(http://corpus.byu.edu/bnc/x.asp)

Man kann auf Basis dieses und ähnlicher Ergebnisse argumentieren, dass die Korpusbeispiele den Sprachgebrauch besser wiedergeben. Der Kontext ist hier beliebig variabel – so könnte der Volltext als Leseverstehens-Übung genutzt werden.

Oft werden Korpora für grammatische Zwecke mit Zusatzinformation versehen, das sogenannte Tagging. Auch Tags können grammatische Informationen beinhalten, aber z.B. auch lexikalisch oder syntaktisch nützlich sein. Dabei mussten die Informationen lange per Hand hinzugefügt werden – inzwischen gibt es automatische Tagger, deren Verlässlichkeit auf über 95% gestiegen ist, nachdem die ersten Programme (z.B. Greene und Rubin mit dem Brown-Korpus) nur bei 70% lagen. Daher war der Nutzen anfänglich nur begrenzt, da beträchtliche Zeit in die Korrektur der automatisch erstellen Ergebnisse investiert werden musste.

Ein solcher auf dem modifizierten Markov-Modell der dynamischen Programmierung nach de Rose/Church basierende Autotagger ist beispielsweise das an der Universität Lancaster entwickelte CLAWS. Auch hier geht der Trend hin zu einer Online-Verfügbarkeit:

Abbildung 8: CLAWS Autotagger

| Originaltext: | WWW-Autotagger mit Tagset c5: |
|---|---|
| "The Federal Reserve would have the power to regulate virtually the entire financial industry under a Treasury Department proposal to be announced Monday." | The_AT0 Federal_AJ0 Reserve_NN1 would_VM0 have_VHI the_AT0 power_NN1 to_TO0 regulate_VVI virtually_AV0 the_AT0 entire_AJ0 financial_AJ0 industry_NN1 under_PRP a_AT0 Treasury_NN1 Department_NN1 proposal_NN1 to_TO0 be_VBI announced_VVN Monday_NP0 ._. |

(http://www.cnn.com/ und http://ucrel.lancs.ac.uk/cgi-bin/claws4.pl)

Oben ein Text vor und nach der automatischen Behandlung mit CLAWS. Man könnte sich vielleicht fragen, was der Vorteil für eine Institution ist, diese Softwarepakete kostenlos verfügbar zu machen. Hier imitieren die Hersteller das erfolgreiche Google-Modell: die Eingabedaten der Onlinebenutzer werden dazu genutzt, die Algorithmen weiter zu verfeinern und statistische Datengrößen zu gewinnen. Nebenher kann das Korpus so leicht erweitert werden. Glücklicherweise ist die Schattenseite dieser Onlinenutzung, nämlich nutzerspezifische Werbung, bislang in der Linguistik noch nicht sehr weit verbreitet.

### 2.2.2 LEXIKOGRAPHIE

Ein zweites zentrales Thema – vielleicht eines der wichtigsten überhaupt für die Korpuslinguistik – ist die Lexikographie. Zur Bestimmung und Untersuchung von Vokabular und Bedeutungen sind große Korpora nötig, in denen die Wörter in einem möglichst breiten Kontext vorkommen, aus dem sich die jeweilige Verwendung ergibt. Mit Hilfe von Konkordanzprogrammen können leicht Wortlisten nach Frequenz oder Kontext erstellt werden. Diese Listen sind sehr hilfreich, wenn es um Begriffsfestlegung geht, und sie sind oft manuell erstellten Bedeutungsfestlegungen überlegen. Der Vorteil frequenzbasierter Wortlisten ist unter Anderem die Möglichkeit, gezielt die am häufigsten verwendeten Begriffe und deren häufigste Bedeutungen aufzulisten. Ein Pionier auf diesem Gebiet war der kürzlich verstorbene John Sinclair, der Gründer des COBUILD-Projektes. Eines der Ziele war die Erstellung korpusbasierter Wörterbücher für Englischlernende, wie z.B. das Collins COBUILD English Dictionary. Sinclair etablierte außerdem eine korpusbasierte Prozedur zur Festlegung der Bedeutung, ausgehend von einem Kernbegriff (der nicht immer alle Korpusbedeutungen abdeckt) über Kollokati-

ons- und Kolligationsmuster hin zu semantischen Feldern und pragmatischer Realisierung. So können Phraseologismen und Kollokationen isoliert werden. Ein Beispiel ist seine korpusbasierte Definition des Begriffs *brink* (2000).

Abbildung 9-1: *brink*

- Definitionen
  - Longman 1995: *to be on the brink*: „to be in a new and very different situation" (Beispielsatz: „Karl is on the brink of a brilliant acting career.")
  - COBUILD 1995: Definition und mehrere Beispielsätze, die den typischen Gebrauch illustrieren, z.B. „Their economy is teetering on the brink of collapse"... „Failure to communicate has brought the two nations to the brink of war."

(Hunston 2002: 101)

Wie man sieht, verwenden korpusbasierte Wörterbücher meist mehrere Beispielsätze, die typische Verwendungsmuster beschreiben. Stellt man den Definitionen einige Sätze aus dem BNC gegenüber, sind diese durch Frequenzanalyse gewonnenen Muster bereits recht gut zu erkennen.

Abbildung 9-2: *brink*

- Auszug aus der BNC-Konkordanz:
  - 83 CK4 , the blustery emotional gales of their live outings still taking me to the **brink** of a seizure. A fact
  - 84 CK6 to Austin, Texas. And despite the fact that Ministry are on the **brink** of a Metallica-scale take-over in
  - 85 EB3 vital 37, lasting a further 24 overs. and seeing England to the **brink** of victory. But his run-out
  - 86 BMF retailers having a difficult time of late. After months of teetering on the **brink**, David Reed has filed
  - 87 C9K less than I thought they knew. But," Jay teeters on the **brink** of a revelation, "the A&R man doesn't

(http://corpus.byu.edu/bnc/x.asp)

Eine Kontextanalyse wie in der nächsten Abbildung isoliert schließlich auch Kollokationen. Die Ergebnisse sind hier nicht nach der jeweils angegebenen Gesamtzahl, sondern nach dem Grad der Mutual Information (MI) geordnet, einer rechnerisch ermittelten Kenngröße für die Wahrscheinlichkeit einer kombinierten Verwendung des untersuchten und des verglichenen Begriffs. Die übliche Verwendung als Kollokation, z.B. *teeter on the brink, poised on the brink, hover on the brink* und der negative Kontext *(bankruptcy, disaster, starvation*) ist klar erkennbar:

Abbildung 9-3: *brink*

- Kontextanalyse im BNC mit Gesamtzahl und MI-Index:

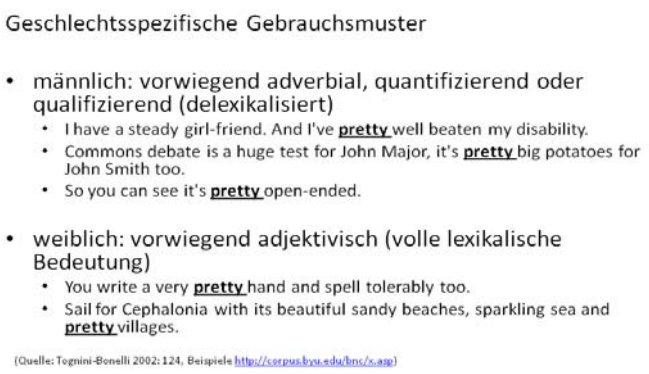
| | | | | | |
|---|---|---|---|---|---|
| 1 | TEETERING 15 | 8.60 | 15 | CIVIL 8 | 3.10 |
| 2 | TEETERED 9 | 8.50 | 16 | PULLED 5 | 2.90 |
| 3 | POISED 11 | 5.97 | 17 | NUCLEAR 6 | 2.85 |
| 4 | STARVATION 6 | 5.73 | 18 | CAREER 5 | 2.74 |
| 5 | HOVERING 5 | 5.71 | 19 | STAND 5 | 2.39 |
| 6 | EXTINCTION 5 | 5.35 | 20 | SUCCESS 6 | 2.36 |
| 7 | BANKRUPTCY 7 | 5.10 | 21 | STOOD 5 | 2.26 |
| 8 | COLLAPSE 17 | 5.06 | 22 | BACK 39 | 2.24 |
| 9 | DISASTER 14 | 4.78 | 23 | ON 259 | 2.15 |
| 10 | DESTRUCTION 5 | 3.93 | | | |
| 11 | REVOLUTION 6 | 3.43 | | | |
| 12 | BROUGHT 21 | 3.22 | | | |
| 13 | WAR 27 | 3.15 | | | |
| 14 | DEATH 19 | 3.11 | | | |

(http://corpus.byu.edu/bnc/x.asp)

Die Verwendung eines Korpus zur Begriffsfestlegung bildet so mit größerer Genauigkeit die Bedeutung eines gegebenen Wortes ab. Fillmore (1992) unterstreicht den Vorteil eines empirischen Wörterbuchs anhand seiner Analyse des Begriffs *risk*. Er findet Beispiele, die nicht unter die Kernbedeutung fallen und somit in vielen Wörterbüchern nicht berücksichtigt werden. Partington (1998: 33ff.) untersucht die Bedeutungen der intensivierenden Adjektiven *sheer, pure, complete, utter, absolute*, die praktisch nur anhand des Korpus zu unterscheiden sind.

Große Korpora ermöglichen auch die Analyse von Schlüsselwörtern und ihres Gebrauchs, so hat beispielsweise Michael Stubbs seine Forschung in lexikalischer Semantik an der Universität Trier auf Korpusdaten aufgebaut und ein korpusbasiertes Dictionary of Keywords in British Culture vorgeschlagen. Wie, wenn nicht durch Korpusdaten, soll beispielsweise der von J.R. Firth (1957: 11ff.) postulierte „kulturelle Rucksack" näher beschrieben werden? Hier als Beispiel der geschlechtsspezifische Gebrauch von *pretty* (nach Aston 2002 in Tognini-Bonelli 2002: 124).

Abbildung 10: *pretty*



Geschlechtsspezifische Gebrauchsmuster

- männlich: vorwiegend adverbial, quantifizierend oder qualifizierend (delexikalisiert)
  - I have a steady girl-friend. And I've **pretty** well beaten my disability.
  - Commons debate is a huge test for John Major, it's **pretty** big potatoes for John Smith too.
  - So you can see it's **pretty** open-ended.
- weiblich: vorwiegend adjektivisch (volle lexikalische Bedeutung)
  - You write a very **pretty** hand and spell tolerably too.
  - Sail for Cephalonia with its beautiful sandy beaches, sparkling sea and **pretty** villages.

(Quelle: Tognini-Bonelli 2002: 124, Beispiele http://corpus.byu.edu/bnc/x.asp)

(Tognini-Bonelli 2002:124, Beispiele: http://corpus.byu.edu/bnc/x.asp)

Dieser Unterschied ist anhand eines geschlechtsspezifischen Nutzerkorpus gut zu identifizieren und zu belegen. Moderne Korpora sind dabei so schnell zu durchsuchen, dass von einigen Linguisten und Linguistinnen sogar die Verwendung von Konkordanzen anstelle des Nachschlagens oder der Begriffsfestlegung in Wörterbüchern empfohlen wird.

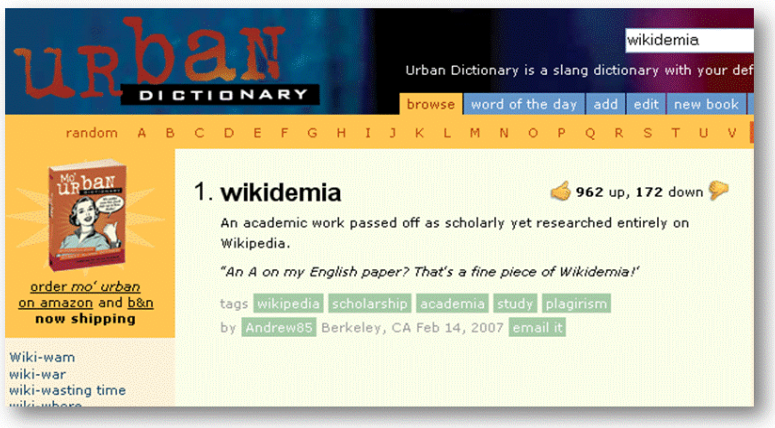
### 2.2.3 LEXIS UND SEMANTIK

Zum dritten Schwerpunkt: Im Bereich Lexis und Semantik gibt es inzwischen eine große Zahl individueller Studien, und die korpusbasierte semantische Analyse ist fast schon zum Standard geworden.

Nur ein Beispiel: Susan Hunston an der Universität von Birmingham, seit John Sinclair ein Schwertpunkt für Korpusforschung, illustriert Bedeutungsmuster und -unterschiede, z.B. negative Nutzungsmuster wie *condemn sth as* (Hunston 2002) Ihre Kollegen Wolfgang Teubert und Anna Cermakova untersuchen 2004 den unterschiedlichen Kontext des Begriffs *Globalisierung* im britischen und amerikanischen Englisch (BrE *globalization*, AmE *globalisation*). Andere Studien befassen sich mit so diversen Themen wie dem Gebrauch des Wortes *reason* in Texten (Hoey 1993) oder den interdisziplinären Varianten des verbalen Hinhaltens (Poos/Simpson, in Reppen 2002: 3-23).

Auch Internetquellen werden zunehmend für die Beschreibung neuer Wörter verwendet. Hier sind vor allem Onlinewörterbücher und Diskussionsforen wie das Urban Dictionary zu nennen. Hier als Beispiel die interessante Neuprägung *wikidemia* für eine auf Wikipedia basierende wissenschaftliche Arbeit (ein Problem, dem man in der Lehre durchaus begegnen kann):

Abbildung 11: Urban Dictionary: *wikidemia*




(http://www.urbandictionary.com)

### 2.2.4 PHRASEOLOGIE

Eng verbunden mit der lexikalischen Forschung ist der Einsatz von Korpora in der Phraseologie. Phraseologische Forschung ist besonders auf authentische Beispiele angewiesen. Kein linguistisches Feld verändert sich so schnell, und In-Phrasen können schon morgen wieder out sein. Darüber hinaus ist traditionell die Beschaffung von empirischen phraseologischen Daten schwierig, zeitaufwändig und sogar – je nach Art und Verbreitung der untersuchten Sprachvariante – gefährlich.

Gerade in den späten 80er und 90er Jahren entstanden zahlreiche Korpora aus Informationsmedien. Mit deren Hilfe können generell verwendete Phraseologismen viel leichter klassifiziert werden. Phraseologismen dienen aber auch zur inneren Abgrenzung von eng begrenzten, oft geschlossenen Gruppen. Diese häufig opaken Begriffe konnten mangels Daten traditionell nur schwer untersucht werden.

In den letzten 10 Jahren hat jedoch praktisch jede phraseologische Variante ein Heim im Internet gefunden, und öffentliche Datenquellen haben die Verständigungsbarriere reduziert. Es ist erstmals möglich, ohne lange Feldforschung z.B. einen Slang von außen zu erfassen und zu untersuchen. Sogar metasprachliche Datenquellen sind häufig. In vielen Internetforen wird über Wortbedeutungen diskutiert. Oft ist die Onlinequelle auch die einzige Möglichkeit, Informationen über innovativen Sprachgebrauch in der Praxis zu erhalten. Dabei variiert natürlich die Qualität der Definitionen beträchtlich:

Abbildung 12-1: Definitionen von *rad*

- Online-Forum:
  - Rad: an abbreviation of 'radical'--a term made popular by the Teenage Mutant Ninja Turtles. […U]sed by people on the West Coast who find words like 'cool', 'awesome', and 'tight' to be tired and overused; 'rad' is generally considered to be a much higher praise […]
    - *"Those are some rad shoes."*
    - *"Oh, RAD."*
  - Rad: it's like saying über cool.
    - *talk like a skater from the 80's, say rad.*

(http://en.wikipedia.org/wiki/Rad)

Das Beispiel zeigt eine sehr differenzierte und eine etwas weniger ausgearbeitete Abgrenzung des Adjektivs *rad* (etwa: ‚toll', ‚cool') im Vergleich mit der Wörterbuch-Definition. Aus den verschiedenen Beispielen ergibt sich aber meist ein gutes Nutzungsmuster, das oft aktueller und anwendungsbezogener ist als die üblichen Beschreibungen. Dies soll hier im nächsten Beispiel demonstriert werden, *schlep*.

Abbildung 12-2: Verschiedene Wortbedeutungen*: (to) schlep*

1. schlep
   - To carry, bring or otherwise transmit something which is difficult to move.
   1. *I'm not going to schlep that stack of papers all the way over here!*
2. schlep
   - A long and tiresome walk. Origin: Yiddish.
   1. *It's a bit of a schlep from here to the shop.*
3. schlep
   - 1. <verb> A very unpleasant or inconvenient journey of any distance.
   - 2. <noun> A person lacking social skills. A dork. Not a friend of choice.
   1. *"Oy, my legs are killin me, I don't wanna schlep to 711 right now!"*
   2. *"Yeah, he always gets blackout drunk! What a schlep."*
4. schlep
   - A person who is completely of low class standards.
   1. *Those teachers are such schleps.*
   2. *Those schleps tried to make me smoke pot.*

(http://www.urbandictionary.com)

Durch den Reichtum an verfügbaren Beispielen und die wachsende mediale Kompetenz der Studierenden ist heute phraseologische Forschung selbst in einführenden Fachseminaren möglich und lohnend. In einem Seminar zur Phraseologie erstellten meine Studierenden beispielsweise Analysen in den Bereichen Hackersprache und Slang auf der Grundlage von Onlinequellen, führten eine Abgrenzung distinktiver Phraseologismen in britischen Tageszeitungen durch, erarbeiteten ein Lexikon der In-Sprache für ERASMUS-Studierende und eines über die Sprache von EU-Bürokraten.

### 2.2.5 KOGNITIVE LINGUISTIK

In der kognitiven Linguistik sind Korpora für die Begriffskonnotation interessant – George Lakoff nutzt in seiner jüngeren Forschung zu Framing (der Bildung von Konzeptstrukturen) nicht nur WWW-Quellen, sondern unterhält auch selbst ein linguistisches Blog (vgl. Lakoff 2004). Einige Beispiele für framing aus der US-Innenpolitik:

Abbildung 13: Beispiele für framing

Begriffe aus der US-Innenpolitik:

- *Tax Relief*
- *No Child Left Behind Act*
- *Pro-life*
- *Pro-choice*
- *Patriot Act*

(http://www.rockridgeinstitute.org, http://cnn.com)

In diese Richtung geht auch die Korpusforschung im Bereich Words in Culture. Die Theorie, dass Texte verbreitete Diskursmuster oder Vorlagen realisieren (vgl. Stubbs 2001, Benveniste 1954/1973) resultierte in zahlreichen Schlüsselwortlisten (z.B. Williams 1983). Ein Korpus kann so Assoziationen aufzeigen, wie ‚wir' gegen ‚sie' (vgl. Partington 1998) und Schlüsselwörter wie *heritage, care* oder *community* beschreiben. Verschiedene Onlinequellen sind in den letzten Jahren dazu übergegangen, Begriffslisten zu veröffentlichen, die über Wörter des Jahres hinaus auch phraseologische Innovation (der *Googleganger* der American Dialect Society fällt mir hier ein) aufzeigen.

### 2.2.6 ÜBERSETZUNGSWISSENSCHAFT

Die Erstellung und Analyse paralleler Korpora in mehreren Sprachen ist vor allem in der Übersetzungswissenschaft nützlich. Teubert und Cermakova (2004: 114ff.) demonstrieren den Einsatz solcher Parallelkorpora als Übersetzungshilfe. Hier die kontextabhängige Verwendung der Begriffe ‚Gebeine'/‚Gräten'/‚Knochen' verglichen mit ihrer englischen Entsprechung *bones*.

Abbildung 14-1: Einsatz von Parallelkorpora (Bedeutungseinheiten)

Problem der unterschiedlichen Bedeutungseinheiten

- Englisch:
  - *bone* „eines der weißen, harten Gewebeteile, die bei Menschen und anderen Wirbeltieren das Skelett bilden" (NOED)
- Übersetzung ins Deutsche:
  - *Gebeine*: „From time to time the bones were dug up"
  - *Gräten*: „All you have left is fine eating without any bones"
  - *Knochen*: „We expect a person to feel terrible after breaking a bone"

(Teubert u.a. 2007:115f.)

Abbildung 14-2: Einsatz von Parallelkorpora (Prozedur)

- Erstellung eines Parallelkorpus
- Isolierung der Übersetzungsäquivalente
- Bildung von Übersetzungseinheiten mit den jeweiligen Kollokationsprofilen
  - z.B. *bone + trout, salmon, eat, fin, remove* > hohe Wahrscheinlichkeit für *Gräte* > Etablierung von Monosemie
  - z.B. *travail > work / labour*

(Teubert u.a. 2007:115f.)

Mit Parallelkorpora ist eine Festlegung der Bedeutungseinheiten und die Erkennung der spezifischen Nutzungsprofile bedeutungsunabhängig und mit Einschränkungen sogar automatisch möglich.

Neben zunehmend größeren Korpora geht es in der Übersetzungsforschung auch um deren Behandlung durch Hinzufügen zusätzlicher Informationen und die verbesserte Auswertung der Sprachdaten. Am Lehrstuhl für Englische Sprach- und Übersetzungswissenschaft der Universität des Saarlandes (Prof. Erich Steiner) ist beispielsweise das Projekt CroCo angesiedelt, das Übersetzungsphänomene wie Explizierung – das Phänomen, dass übersetzte Texte deutlicher, expliziter als die dazugehörigen Ausgangstexte wirken – untersucht und entsprechende Parallelkorpora annotiert. Das Zentrum für Übersetzung und Interkulturelle Studien der Universität Manchester (Prof. Mona Baker), um ein weiteres Beispiel zu nennen, hat das Translational English Corpus kompiliert und entwickelt spezifische Konkordanzsoftware weiter.

### 2.2.7 HISTORISCHE LINGUISTIK

Eine wichtige Untersuchungsrichtung ist die historische Linguistik. Die diachronische Untersuchung von Veränderungen der Sprache, wie Änderungen in Lexis und Syntax, wird durch Korpora erheblich erleichtert. Sowohl Erstvorkommen (z.B. Wegfall von Flexionsformen) als auch Trends (z.B. Gebrauch von Fremdwörtern) können dokumentiert werden. Das geht umso besser, je größer die verfügbaren Korpora sind. Diachronische Korpora werden so in der historischen Linguistik zu einem immer wichtigeren Hilfsmittel, zumal die technischen Probleme bei der Korpuserstellung geringer werden. Nur ein Beispiel: In Diemer (1998) werden alt- und mittelenglische Korpora verwendet, um frühe Belege für die Verwendung des Präfixes *out-* in der Bedeutung ‚jemanden übertrumpfen' zu finden.

Abbildung 15-1: *out-* im Altenglischen

Erweiterte direktionale Bedeutung, z.B. *fight out (of)*

- *Gif hit +tonne hwa do, +donne sie he scyldig cyninges mundbyrde & +t+are cirican fri+des mare, gif he +d+ar mare ofgefo, gif he for hungre libban m+age, buton he self utfeohte.*
  *("es sei denn, er kann sich frei- / herauskämpfen")*

(Helsinki-Korpus colaw2: Laws of King Alfred)

Im Altenglischen ist *out-* als Präfix im Korpus in direktionaler Bedeutung häufig; hier ist wohl der Ursprung der ‚übertrumpfen'-Bedeutung (‚weiter in eine Richtung gehen als jemand Anderes') zu sehen, wie *ut(a)drifan*. Interessant ist *outfight* aus dem Helsinki-Korpus.

Abbildung 15-2: out- im Mittelenglischen

- *to output (hinauswerfen)*
  - *and than outpute thaim fra the heritage of heuen, fere as thaire wickidnes diserues; for thai excitid the til vengaunce, duelland in thaire synne*
- *to outfight* (im Sinne von *überwältigen*):
  - *For ri3twisnesse fi3t for thi soule […] and God shal Outfi3ten, 'or ouer come', thin enemyes for thee*
  - (King James, Ecclesiasticus 4, 33: *Strive for the truth unto death, and the Lord shall fight for thee*; die Vulgata hat: *pro iustitia agoniare pro anima tua et usque ad mortem certa pro iustitia et Deus expugnabit pro te inimicos tuos*)

(Helsinki-Korpus, Wycliffe-Korpus)

Im Mittelenglischen finden sich zunehmend abstrakte Bedeutungen, vor allem in der Wycliffe-Bibel um 1390. Im zweiten Beispiel sieht man schön, wie der Übersetzer hin und her überlegt hat. Die Bedeutung ist also wohl noch ungewohnt.

Weitere Beispiele aus der Forschung sind die historische Veränderung von Partikelverben, Lexikalisierung und Grammatikalisierung. Auch hier können anhand von Korpora zunehmend Trends genauer analysiert und Entwicklungen illustriert werden. Laurel J. Brintons und Elisabeth Closs Traugotts (2005) These eines Lexikalisierungsprozesses bei präpositionalen Verben lässt sich auf der Basis einer Korpusanalyse z.B. nachvollziehen und illustrieren. Gleiches gilt für den von Paul J. Hopper und Closs Traugott (2003) beschriebenen Grammatikalisierungsprozess, also den Verlust lexikalischer Inhalte zugunsten der Übernahme grammatischer Funktionen.

### 2.2.8 DIDAKTIK UND FREMDSPRACHENERWERB

Auch in der Didaktik, insbesondere dem fachsprachlichen Fremdsprachenerwerb, kann der Einsatz von Korpora sinnvoll sein. Korpora ermöglichen beispielsweise ein genaues Studium des Lernfortschritts und der Fehlerverortung bei Sekundär- und Fachsprachen. Viele dieser Möglichkeiten sind auch in der universitären Lehre nutzbar, z.B. auf der Grundlage von Elena Tognini-Bonellis Arbeit (Corpora in ELT, 2001).
Die Erstellung von Lernkorpora erlaubt eine Analyse von typischen Fehlern in Grammatik oder Vokabular. Analog zur phraseologischen Forschung können mittels Schlüs-

selwortanalyse (z.B. bei Scott u.a. 2006) Hauptfehlerquellen gefunden und didaktisch aufbereitet werden.

Die Nutzung realer Sprachbeispiele macht Fremdsprachenlehre wesentlich authentischer. Der Einsatz von Korpora führt auch zu didaktischer Innovation und in der Folge zu autonomerem Lernen und erhöhter Motivation. Beim datengetriebenen Lernen oder Discovery Learning nutzen die Lernenden deduktive anstelle von induktiven Methoden und entdecken so vorhandene Muster mit Hilfe eines didaktischen Korpus. Bernardini (2004:23f.) beschreibt beispielsweise die Analyse von komplexen Zeitungsüberschriften.

Lernkorpora wie CHILDES oder das International Corpus of Learner English können als Parallelkorpora zur Fehleranalyse genutzt werden. Im Übersetzungslernbereich können so typische Übersetzungsfehler und false friends aufgezeigt werden (*seriös* > *serious*).

Aber auch für die Lehrenden ist der Einsatz von Korpora interessant. Bei der Vorbereitung ersetzt, wie Dieter Mindt (1995) beobachtet, aktueller Sprachgebrauch so Introspektion und Intuition, z.B. bei der Erklärung grammatischer Sachverhalte wie Zeitenfolge oder Gebrauch des Simple Present mit den Aspekten: ‚aktuell', ‚gewöhnlich', ‚ausgedehnt'. Die Lehrenden können Unterschiede besser illustrieren, gerade bei fast synonym verwendeten Begriffen wie *day by day*, *day after day* oder *tall, high* (Partington 1998, vgl. auch Tsui 2004:44ff.). Lehrenden-Korpora und Chats wie TeleNex (ibid.) ermitteln Diskrepanzen zwischen Lehrbuch und Sprachgebrauch und bereiten sie didaktisch auf.

### 2.2.9 SOZIOLINGUISTIK

Die soziale Variation von Sprache ist seit jeher ein attraktives Forschungsgebiet. Besonders William Labov und seine Methode der Datensammlung im Rahmen seiner Studien zur Stratifikation in New York City (1966) hat die Korpusforschung in diesem Bereich geprägt. Dabei stellt sich allerdings immer das Problem des Beobachterparadoxes und natürlich das der Logistik der Datenerhebung in Umfragen, ganz abgesehen von zeitlichen und finanziellen Einschränkungen.

Diese Probleme sind heute wesentlich geringer geworden. Auf Webportalen und Communities wie YouTube und Facebook stellen Nutzende heute freiwillig linguistische Daten zur Verfügung, von Erzählungen über Fachtexte bis hin zu persönlichen Reflexionen. Wir mögen erstaunt sein über diese Bereitschaft, Persönliches öffentlich zu machen, aber wir erhalten so eine ganz neue Qualität von Korpora und eliminieren das alte Problem der Beeinflussung der Ergebnisse.

Insbesondere bei der Untersuchung der verschiedenen fachlichen Sprachvarianten ist der Einsatz von Korpora sinnvoll. Hier sind sowohl historische Varianten (Rechts- und medizinische Sprache) als auch moderne Ausprägungen (Wirtschaftsenglisch, Hacker Language, Political Correctness) zu nennen. Fachsprachen sind heute häufig durch ironisierende, individuelle Wortschöpfungen, Kurzlebigkeit und einen noch höheren Grad der Opazität geprägt. Ohne entsprechende Foren wäre es fast unmöglich, diese Sprachvarianten zu beschreiben:

Abbildung 16: Hacker Language

- Person 1: "Hey man, how's it going"
  Person 2: "Meh"
  Person 1: "I started a dA"
  Person 2: "SKOC?„
  Person 1: "Yeah, lol. How do I use DOSBox?"
  Person 2: "Uhh, Line1: Mount <drive> <directory(inc
  drive)> Line2: <drive> Line3: CD <Game folder>"
  Person 1: "Thanks, g2g bai"
  Person 2: "Bai lol„
- Leet: l337
  Skills: s|<1llz

(http://www.urbandictionary.com)

Dieses Beispiel aus einer Unterhaltung zwischen zwei Computer-Hackern wäre beispielsweise ohne Erläuterung kaum verständlich, ist aber hochinteressant für die Analyse. Auch das zweite Beispiel, der sogenannte Leet-Code, ist fast völlig opak. Bei der Etablierung von sozialen Trends sind Onlinequellen ebenfalls hilfreich. So untersucht Stenström (2002) Ausbreitungsmuster in der Teenager-Sprache. Oft kommt es zu innovativen Verknüpfungen und Imitationen von Konstruktionen aus anderen Sprachen (im Marketing z.B. *old-old* für *very old* etc.). Dieser Trend geht einher mit der Entwicklung zu immer spezielleren Fachsprachen (Fan-Sprachen, Community-Sprachen). Dabei weicht die Websprache immer mehr von der Schriftsprache ab. Sie ist zunehmend interaktiv, oft in Echtzeit, durch Hypertext ergänzbar und verknüpft mit Zusatzinformationen (Community Websites, verknüpfte Fotos). In der Syntax ist einerseits ein Trend zu zunehmender Komplexität und andererseits eine teils extreme Vereinfachung (Jugendsprache, Drogenslang) zu beobachten. Hier einige schöne Beispiele für Vereinfachungen, die aus der Nutzung von Textbotschaften und Chats resultieren.

Abbildung 17: Abweichungen vom Standard

*"How are you guys?"*
*> "Hwo 4r3 you gys"*
*> "h0w r u?"*
*> "u"*

- *Lol* (in Dialogen)
- *could of*
- *BCNU*
- *Mine.* (für *„das gehört mir"*)

(Google-Korpus, http://www.urbandictionary.com)

In Onlinequellen treten außerdem zunehmend Fehler auf, die unter Umständen die ersten Zeichen für neue Sprachentrends sind. Hier bietet das Internet ungeahnte Multiplikatoren für zuvor idiosynkratische Formen. Ist beispielsweise der Begriff *swordman* statt *swordsman* noch ein Fehler? Im Google Corpus stehen immerhin schon 900 Vorkommen von *swordman* 3 700 Vorkommen von *swordsman* gegenüber.

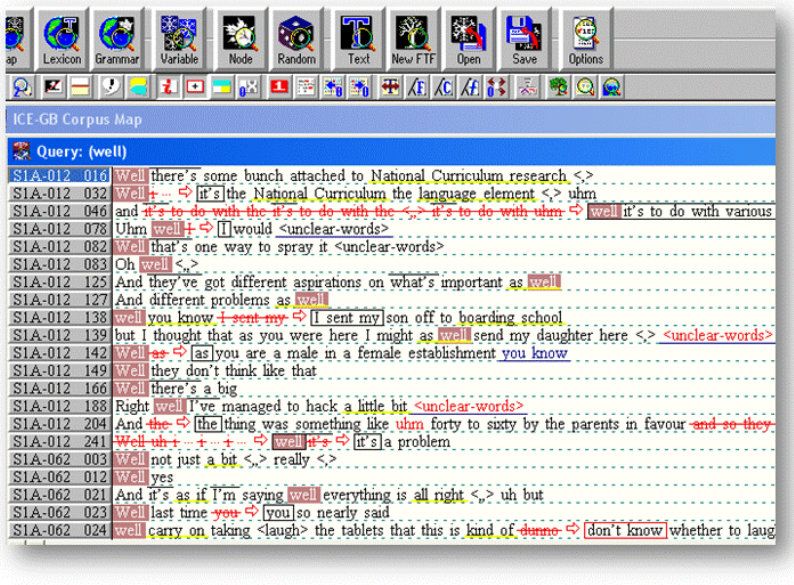
Soziale Varianten und Fachsprachen bieten eine Schnittstelle der Korpuslinguistik zur angewandten industriellen Forschung, z.B. bei der Erstellung von Online-Lehrmodulen oder der Erstellung von Frequenzlisten für Übersetzungsprogramme.

### 2.2.10 GESPROCHENE SPRACHE

Lange Zeit war es sehr schwierig, die gesprochene Sprache mit Hilfe von Korpora zu untersuchen, z.B. in der Gesprächs- oder Diskursanalyse oder der Phonologie.

In den letzten 10 Jahren sind aber sowohl in der geplanten Rede (vor allem Fernseh- und Rundfunknachrichten) und auch im spontanen Dialog (Transkriptionen von Erzählungen, Scherzen, Telefongesprächen) Korpora transkribiert und annotiert worden. So können nicht nur Inhalt, sondern auch prosodische Elemente, Überlappung und Diskursmarker untersucht werden. Die Korpusforschung ermöglicht so auch in diesem Bereich inzwischen die Erforschung von Einzelphänomenen mit ausreichender Datengrundlage. Hier ein Ausschnitt aus dem ICE-GB Korpus des University College London, der die Verwendung des Diskursmarkers *well* untersucht.

Abbildung 18: Gesprächsanalyse




(ICE-GB Sample Corpus)

### 2.2.11 SPRACHVERARBEITUNG / COMPUTERLINGUISTIK

Als letztes Beispiel soll kurz auf den Nutzen von Korpora in der Sprachverarbeitung und Computerlinguistik eingegangen werden. Hier konzentrieren sich zurzeit ein Großteil der verfügbaren Drittmittel und die Forschung in universitären wie außeruniversitären Einrichtungen wegen der naheliegenden praktischen Anwendungen, die hauptsächlich in den Bereichen Spracherkennung und Datenbankoptimierung liegen. Die Themen sind dabei vielfältig und reichen von der Verknüpfung mehrsprachiger Datenquellen im europäischen Kontext (vgl. z.B. das Projekt „Intera" des Institut für Computerlinguistik an der Universität des Saarlandes) bis hin zu Erstellung von Programmen zur sprachgesteuerten Interaktion von Mensch und Maschine.

Korpora spielen in diesem Bereich eine entscheidende Rolle. Für die Forschung sind möglichst große aktuelle Textdatenmengen nötig, damit die Algorithmen etwa für Spracherkennung oder Informationsverknüpfung eine größere Vergleichsbasis haben. Je größer die Datengrundlage, desto genauer die Ergebnisse bzw. desto besser die Funktion des entsprechenden Programms. Der Hauptgrund für die Entwicklung des vorhin erwähnten Google Korpus liegt beispielsweise nicht in der Sprachforschung, sondern in der Optimierung von Suchvorgängen. Es ist abzusehen, dass dieses Forschungsfeld mit der zunehmenden Rolle interaktiver Computeranwendungen noch an Wichtigkeit gewinnen wird.

## 2.3 GRENZEN DER KORPUSLINGUISTIK

Trotz der beschriebenen Vorteile sind Korpora natürlich noch nicht perfekt. Was sind also die Grenzen der Korpuslinguistik?

Typische Fehlerquellen bei der diachronischen Forschung sind eine zu geringe Datenmenge und die Schwierigkeit der orthographischen Variation. Bis heute sind die existierenden historischen Korpora unvollständig, und viele Studien basieren noch auf den Daten des Helsinki-Korpus. Das liegt hauptsächlich an der Schwierigkeit der Übertragung in elektronisches Format. In der synchronischen Sprachwissenschaft liegt die Problematik eher im Vorhandensein zu vieler ungefilterter Daten, deren statistische Relevanz schwierig festzulegen ist. Aber es gibt auch grundlegende Einwände. Michael Stubbs nennt 2001 zwei Grundprinzipien von Korpusstudien:

Erstens: Korpuslinguistik ist eine empirische Wissenschaft. Als solche muss sichergestellt werden, dass die beobachteten Daten nicht beeinflusst werden. In diesem Zusammenhang ist es oft schwierig, ein geeignetes Korpus festzulegen, das nicht durch die Fragestellung selbst vorselektiert ist.

Zweitens: Wiederholte Ereignisse sind relevant (cf. Stubbs 2001: 221). Dieses Prinzip wird regelmäßig von der theoretischen Linguistik in Frage gestellt, da je nach Untersuchungsgegenstand auch Einzelvorkommen den gewünschten Nachweis liefern können. Er fügt sechs Hauptkritikpunkte hinzu:

*(1) Korpora sind nicht repräsentativ.*

*(2) Korpora liefern nur positive Daten.*

*(3) Bei Korpusdaten fehlt oft der Kontext.*

*(4) Einzelvorkommen werden überbewertet.*

*(5) In Korpusstudien geht Variation durch Nivellierung verloren.*

*(6) Korpusstudien untersuchen Performanz und nicht Kompetenz.*

Ohne auf alle Punkte im Einzelnen einzugehen, ist festzuhalten, dass die meisten dieser Probleme durch den technischen Fortschritt der letzten Jahre gelindert wurden. Korpora wie das BNC sind inzwischen so groß, dass selbst einzelne Diskursarten repräsentativ sind. Gleichzeitig ist die Analyse so differenziert, dass weder Einzelvorkommen untergehen noch der Kontext verloren wird. Was den letzten Punkt angeht (Performanz statt Kompetenz), handelt es sich im Wesentlichen um Chomskys Einwand. Mit den zunehmend vorhandenen Lernkorpora kann allerdings aus mangelnder Performanz durchaus auf die zugrundeliegende Kompetenz geschlossen werden. Viel wichtiger als diese eher prozedurbezogenen Kritikpunkte scheinen mir zwei Einwände:

Erstens: Korpuslinguistik liefert keine Interpretationen, sondern nur unterstützende Beweise. Durch übermäßige Interpretation kann auf der Grundlage zu geringer Daten ein völlig falsches Bild gezeichnet werden. Hier erscheint es sinnvoll, primär deskriptiv zu arbeiten und erst im Anschluss behutsam zu interpretieren, falls nötig.

Zweitens: ein verlässliches Referenzkorpus, ein „Corpus of English(es)", das die ganze Sprache in ausreichender Größe abbildet, existiert noch nicht. Die Gegenwartssprache entwickelt sich ständig weiter. Hier wird die Aufmerksamkeit der Forschung sich immer mehr auf das Internet richten, das in seiner Gesamtheit erstmals ein Studium von Sprache im großen Maßstab und so etwas wie Echtzeit möglich machen könnte.

Wie weit ist diese unvergleichliche Ressource schon nutzbar, und wie wird die Zukunft der Korpuslinguistik aussehen? Darauf soll im letzten Punkt eingegangen werden.

## 2.4 DIE ZUKUNFT DER KORPUSLINGUISTIK

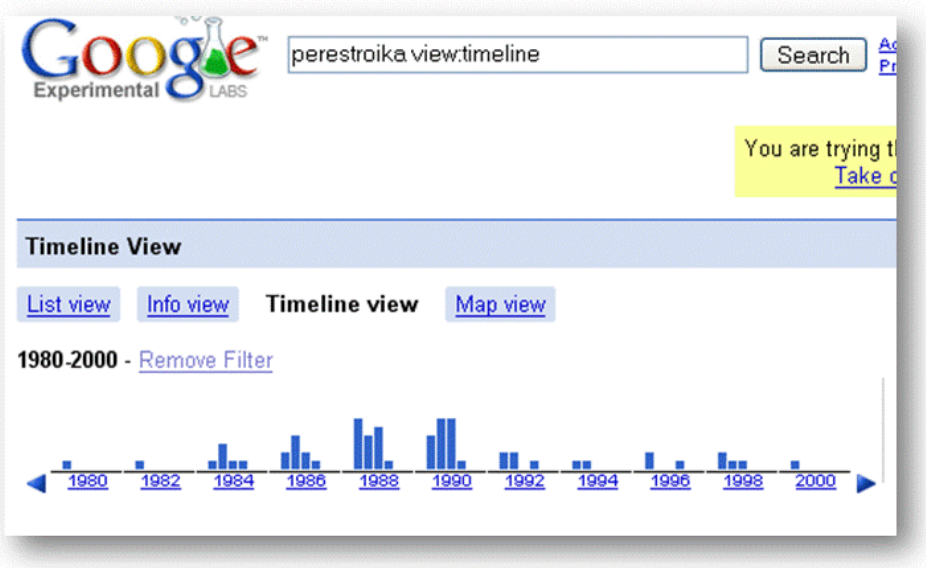
In einer berühmten Fernsehserie der 1980er stellen Raumfahrer aus einer fernen Zukunft einem Supercomputer die Frage nach dem Sinn des Lebens. Eine solche Frage einer künstlichen Intelligenz zu stellen, erscheint den meisten von uns absurd. Es ist daher einigermaßen verwunderlich, dass „Wer ist Gott?" 2007 eine der häufigsten an das Suchportal Google gestellten Fragen war (Google Zeitgeist, http://www.google.com/intl/en/press/zeitgeist2007/mind.html). Die Qualität der Antworten ist erstaunlicherweise hoch. Die Fragesteller nutzen hier natürlich kein „Superhirn", sondern sie rufen inhaltliche Informationen aus einer virtuellen, von Nutzern erstellten Datenmenge. Es ist – zumindest für viele Leute unter 40 – durchaus selbstverständlich geworden, Portalen wie Google oder Wikipedia online Fragen wie „Karte New York", „berühmte Linguisten" oder „Job TU Berlin" zu stellen und das eigene Handeln nach den Antworten auszurichten.

Diese Tendenz geht auch an der Linguistik nicht vorbei. Liegt also die Zukunft des Korpus im Internet? Google oder andere Suchmaschinen erscheinen zunehmend als die Möglichkeit des ultimativen Korpusanalysetools, bei dem die Interpretation gleich mitgeliefert wird.

Unlängst hat Google eine Reihe experimenteller Suchmaschinen entwickelt, die, wenn sie über das Versuchsstadium hinaus sind, die linguistische Korpusforschung revolutionieren könnten. Im Folgenden möchte ich einige Beispiele für den linguistischen Nutzen solcher Entwicklungen zeigen:

Mit einer sogenannten Timeline-Suche können im Google-Portal online Begriffe mit in Dokumenten erhaltenen Daten verknüpft werden. Der Begriff *perestroika* zeigt beispielsweise die größte Häufigkeit ab 1984 bis 1992. Eine auf dem „klassischen" BNC basierende Analyse liefert ein ähnliches Ergebnis und ist darüber hinaus weit schwieriger zu implementieren.

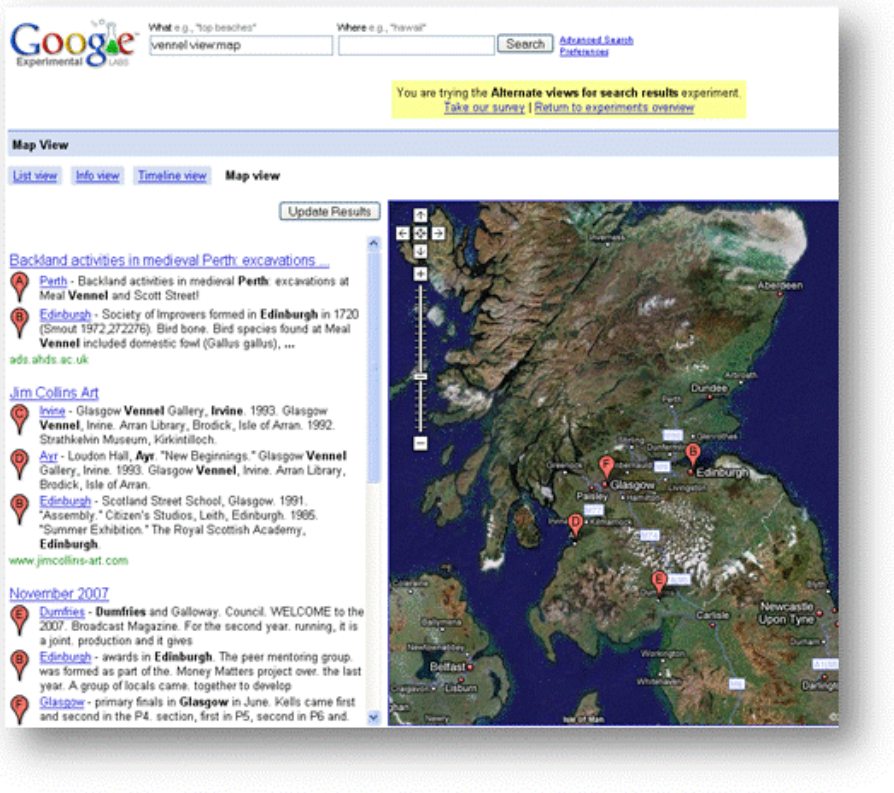
Abbildung 19: Diachronische Suche




(Google Labs, http://www.google.com/experimental)

Mit einer geographischen Suche kann der Ursprungsort von Dokumenten mit darin enthaltenen Begriffen verknüpft werden. Dies ist natürlich für die Dialekt- und Phraseologieforschung von ungeheurem Interesse. So könnte eine bestimmte Schreibweise oder ein dialektaler Begriff verortet werden, wie z.B. *vennel* (schottisch für Gasse) – aber vorerst nur, wenn der Begriff nicht mehrdeutig ist.

Abbildung 20: Geographische Suche

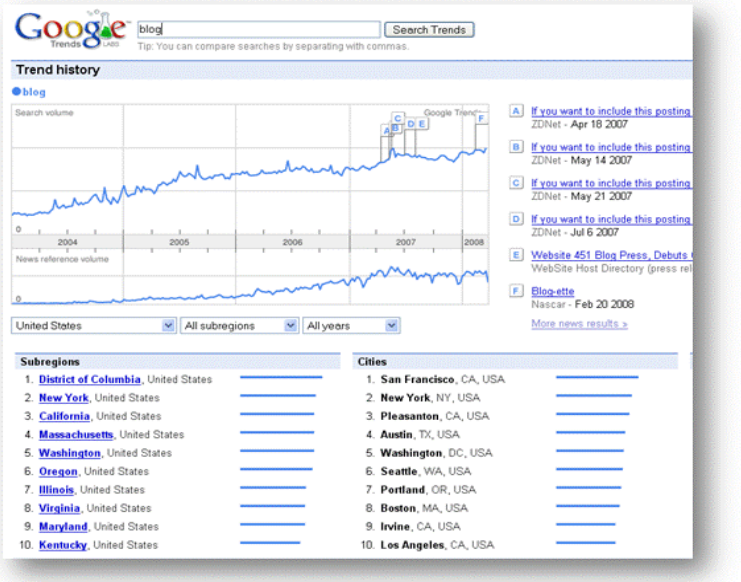



(Google Labs, http://www.google.com/experimental)

Das statistische Analysetool Google Trends untersucht die Verwendung von Begriffen durch die Häufigkeit der Suchabfragen. So kann die Beliebtheit neuer Phraseologismen und lexikalischer Innovationen belegt werden.

Eine Suche nach dem Begriff *blog* (verkürzt für *Weblog*) illustriert im Folgenden die wachsende Beliebtheit des Begriffs (und der Kommunikationsform). Eine traditionelle diachronische Analyse verschiedener nach statistischen Gesichtspunkten ausgewählten Publikationen und eine entsprechende quantitative Analyse würde mit weit größerem Aufwand ein vergleichbares Bild liefern. Dabei bietet das Programm sogar die Einschränkung der Suche auf Staaten und sogar auf Städte an. Es wäre so leicht möglich, das erste Auftreten eines Begriffs wesentlich genauer zu verorten, als das mit traditionellen Methoden der soziolinguistischen Feldforschung möglich war. Im Beispiel ist (stark vereinfacht) eine stärkere Verwendung des Begriffs *blog* in städtischen Gebieten als in ländlichen zu beobachten.

Abbildung 21: Trend



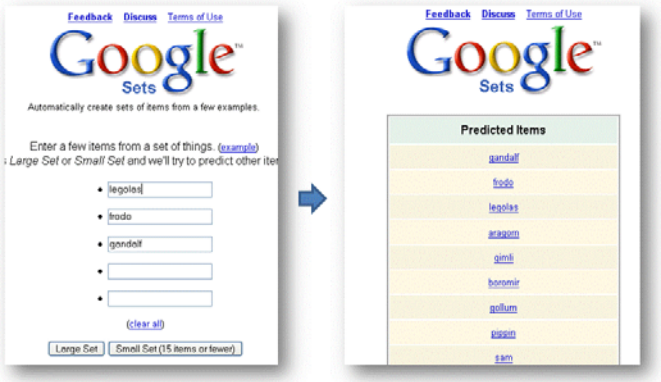

(Google Labs, http://www.google.com/trends)

Praktisch für kognitive Linguisten und Linguistinnen maßgeschneidert ist der von Google Sets verwendete Algorithmus. Hier werden beliebige Suchbegriffe in ihren Textumgebungen quantifiziert und mit den Wörtern, mit denen sie am häufigsten auftreten, aufgelistet. Auch verschiedene Wortsuchen, die auf die gleiche Internetseite führen, werden analysiert. Auf diese Weise werden relationale Daten erstellt, die die Wahrscheinlichkeit beschreiben, dass Begriffe gemeinsam auftreten. So können nicht nur Idiome, sondern auch länger Listen gefunden werden, die Google ‚Sets' nennt. Dabei werden die Ergebnisse genauer, je mehr Bestandteile des Sets bekannt sind.

Ein entscheidender Vorteil des WWW gegenüber einem traditionellen Korpus ist hier vor allem die Größe des untersuchten Texts. Aber selbst mit dem relativ großen British National Corpus und unter Verwendung existierender Korpusanalyse-Programme wäre es schwierig, die Kontextanalyse so massiv zu erweitern, dass alle Kontextbegriffe erfasst werden könnten.

Ein einfaches Beispiel kann das demonstrieren: Eine Suche nach Legolas, Frodo und Gandalf, drei der Hauptfiguren aus J.R.R. Tolkiens „Herr der Ringe", liefert als Ergebnis die komplette Liste der neun Gefährten aus dem Roman. Wird nur ein Name eingegeben, ist die Liste weniger spezifisch und bezieht auch andere Charaktere ein. Diese automatische Verknüpfung thematisch verwandter Begriffe ist mit traditionellen Analysemethoden nicht oder nur sehr schwer möglich. Die meisten Konkordanz-Programme begrenzen den Abstand zwischen Suchbegriff und Kontext auf höchstens 15 Wörter nach beiden Seiten im laufenden Text. Im Unterschied dazu ermöglicht Google Sets die Berücksichtigung z.B. eines kompletten literarischen Werks in einem einzigen Suchvorgang.

Abbildung 22-1: Sets: LoR




(Google Labs, http://labs.google.com/sets)

In der linguistischen Forschung ist dieser Algorithmus sehr hilfreich für grundlegende Fragen des Framings, also des vorausgesetzten Kontexts. Eine Set-Erstellung für *British* und *heritage*, nachfolgend abgebildet, zeigt beispielsweise deutlich deren Konnotation. Die Liste führt *English garden und English home, realm, majesty* und *ancestry* auf, und hilft, den Begriff *heritage* im britischen Kontext wesentlich genauer linguistisch zu beschreiben. Weitere Anwendungsmöglichkeiten liegen in der Untersuchung von Kohäsion und Syntax, der Phraseologismen-Forschung (von Verbverbänden bis zu Sprichwörtern) oder der kognitiven Linguistik. Die Option der Erweiterung des Sets auf mehr als 15 Begriffe ermöglicht sogar die Untersuchung kompletter semantischer Felder (wenn auch auf einer unklaren statistischen Basis, wie viele der vorgestellten Programme).

Abbildung 22-2: Sets: *British heritage*


(Google Labs, http://labs.google.com/sets)

Zuletzt ein Algorithmus mit dem interessanten Namen Google Zeitgeist. Hier werden Begriffe gesucht, deren Onlinenutzung besonders schnell zu- bzw. wieder abnimmt was auf Englisch *zeitgeisty* genannt wird. Dabei ist auch eine Differenzierung nach der Art des Anstiegs und der Zahl der Suchen innerhalb eines bestimmten Zeitraums möglich. Im unten abgebildeten Beispiel werden *fast gainers* für das Jahr 2008 aufgeführt, also Begriffe, die besonders schnell in das Suchverhalten eingedrungen sind.

Die Ergebnisse können nach Ländern, Regionen und Städten getrennt werden, was für die Soziolinguistik und die Sozialwissenschaften interessant ist. Im Beispiel wurde eine Differenzierung in USA und die Welt insgesamt vorgenommen. Für 2008 finden sich so *iphone* und *youtube*. Eine Unterteilung in Quartale ergibt mit *Anna Nicole Smith, iphone, Hurrican Dean* und *Pavarotti* ein exaktes Abbild der beliebtesten Jahresthemen. Mögliche linguistische Anwendungen liegen in der Lexikologie, z.B. bei der exakteren Einschätzung lexikalisch relevanter Begriffe (*fast gainers* sind offensichtliche Wörterbuch-Kandidaten), in der Beschreibung aktueller soziolinguistischer Trends oder in der Differenzierung zwischen Regionen und Ländern. Die außerdem von Google angebotenen Rubriken *Newsmakers* (für Namen), *Showbiz, All the Rage* (für Konsumartikel) und *Top of Mind* (für abstrakte Begriffe und Fragen) lassen nach Abschluss der Testphase weitere Anwendungsmöglichkeiten erwarten.

Nach so viel Hype folgt nun der Haken: Natürlich ist es problematisch wenn Studierende wie Forschende Onlinedaten zunehmend instinktiv sowohl im theoretischen Bereich als auch für datenbasierte Forschung nutzen. Es gibt keine klare Grundmenge, und die quantitativen Daten sind, falls überhaupt relevant, lediglich relativ zu bewerten. Eine Orientierung an diesen Ergebnissen kann zu gravierende Interpretationsfehlern führen. Es sollte auch nicht vergessen werden, dass eine gewinnorientierte Firma wie Google durchaus die Resultate verfälschen kann und das auch tut – beispielsweise werden je nach Region bestimmte Suchbegriffe herausgefiltert. Die Ansätze für wissenschaftliche Suchmaschinen, wie z.B. WebCorp, bleiben allerdings weit hinter der Leistung von Googles PageRank-Algorithmus zurück. Es ist für öffentliche Forschungseinrichtungen aus praktischen und finanziellen Erwägungen auch kaum möglich, den großen Portalen Konkurrenz zu machen. Ergebnisse wie die oben vorgestellten sind zwar vorerst noch nicht wirklich linguistisch nutzbar, doch der Vergleich mit Resultaten aus dem BNC und anderen Referenzkorpora (soweit möglich) liefert oft ähnliche Ergebnisse. Das Potenzial ist jedenfalls beachtlich.

## 3 ZUKÜNFTIGE ENTWICKLUNGEN

Wie sollte also dieses noch unvollkommene Medium genutzt werden? Meiner Meinung nach hat die Entwicklung eines Analysetools zur statistischen Erfassung von Vorkommen innerhalb offener Grundmengen klare Priorität. Niemand weiß genau, wie groß das Web ist, aber bei genügend großen Mengen ist es statistisch ohne Weiteres möglich, mit einer offenen Grundmenge zu arbeiten. So könnte die Relevanz von webbasierten Ergebnissen gesteigert werden. Diese Kooperation ist bereits an mehreren Forschungszentren (nicht nur im Google-Labor) im Gange, z.B. in der Computerlinguistik an der Universität des Saarlandes, oder natürlich die klassischen Zentren der Korpusforschung, z.B. an der Birmingham City University, wo von Jayeeta Bannerjee das bereits erwähnte WebCorp entwickelt wurde. Weiterhin sollten bereits existierende Korpora

mit den neuen, webbasierten Methoden vernetzt werden, etwa durch Integration in Online-Datenbanken. Ein ganz wesentliches Problem ist hier aber die urheberrechtliche Zugriffsbeschränkung vieler Korpora. In den Bereichen Lexis, Phraseologie und Soziolinguistik sind die webbasierten Quellen ausgereift genug, um Sprachwandel zu dokumentieren. Insbesondere die mit dem Internet verbundenen Kommunikationsmöglichkeiten (z.B. Foren, Mailserver, Chats) bieten eine neue Qualität, Sprache zu prägen. Die Rolle dieser Medien und die Beschaffenheit des medialen Diskurses sind für die angewandte linguistische Forschung von außerordentlichem Interesse. Bereits vor mehr als 15 Jahren beschäftigte sich David Crystal mit Internet Language. Er war zu dieser Zeit der Auffassung, dass das neue Medium eine linguistische Revolution einleiten würde. Und das war vor Blogs, Chats und Internet Communities. Wer internetbasierte Sprache untersucht, kann durchaus zu dem Schluss kommen, dass hier eine Gruppe von Sprachvarianten entstanden ist, die sich zwischen den alten Dichotomien von Schrift- und gesprochener Sprache, von Hoch- und Gemeinsprache, bewegen.

Wie wird sich die zukünftige Korpusforschung entwickeln? Besonders interessant sind die bereits geschilderten Möglichkeiten des virtuellen Korpus. Mit entsprechenden Algorithmen und bei ausreichend großer Datenmenge sollte es möglich sein, verlässliche Ergebnisse selbst für sehr spezifische linguistische Fragestellungen zu erzielen.

Doch Korpustechnologie ist kein Selbstzweck, auch wenn der Nutzen von klaren und statistisch zu belegenden Ergebnissen für das Einwerben von Drittmitteln förderlich ist. Aber auch – und ganz besonders – die Nutzung von Korpora für die Hochschuldidaktik und zur Förderung wissenschaftlicher Nachwuchsforschung ist eine lohnenswerte Aufgabe. David Crystal sagte im Jahr 2004:

> Die linguistische Originalität und Neuheit des Internets sollte unsere Herzen schneller schlagen lassen. Es bietet uns eine Zukunft, in der unsere Kommunikation sich radikal von der der Vergangenheit unterscheiden wird. [...] Es ist aufregend, von Anfang an dabei zu sein. (Crystal 2004: 35)

Die korpuslinguistische Forschung ist heute ideal platziert, um Sprache und Sprachwandel zu beschreiben und zu interpretieren. Die Entwicklung in den kommenden Jahren wird zeigen, welche Rolle das neue Medium Internet hier spielen kann.

## LITERATUR

Aston, Guy & Lou Burnard. 1998. *The BNC handbook: exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.

Baker, Mona. 2009. *Critical Readings in Translation Studies*. London: Routledge.

Banerjee, Jayeeta, Antoinette Renouf & Andrew Kehoe. 2006. WebCorp: An integrated system for web text search. In Marianne Hundt, Nadja Nesselhauf & Carolin Biewer (Hrsg.), *Corpus linguistics and the web*, 47-68. Amsterdam: Rodopi.

Benveniste, Émile. 1973. *Problems in general linguistics*. Miami: Miami University Press.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan (Hrsg.). 2004. *Longman grammar of spoken and written English*, 4. Aufl. Harlow: Longman.

Chomsky, Noam. 1995. *The minimalist program*. Cambridge, Massachusetts: MIT Press.

Coates, Jennifer. 1983. *The Semantics of the modal auxiliaries*. London: Croom Helm.

Crystal, David. 2004. Oh what a tangled web we weave. *Science & Spirit* Nov.-Dec. 34-35

Davidse, Kristin, Tine Breban & An Van linden. 2008. The development of secondary deictic meanings by adjectives in the English NP. *English Language and Linguistics 12 (3)*. 475-503.

Davies, Mark. 2006. *A frequency dictionary of Spanish: Core vocabulary for learners*. London: Routledge.

Diemer, Stefan. 1998. *John Wycliffe und seine Rolle bei der Entstehung der englischen Schriftsprache*. Frankfurt: Lang.

Fillmore, Charles. 1992. Corpus linguistics or computer-aided armchair linguistics. In Jan Svartvik & Randolph Quirk (Hrsg.), *Directions in corpus linguistics*, 35-60. Berlin: Mouton de Gruyter.

Firth, John R. 1957. *Papers in Linguistics 1934-1951*. London: Oxford University Press.

Granger, Sylvianne. 1998. *Learner English on computer*. London: Longman.

Grimm, Friedrich. 1891. *Der syntactische Gebrauch der Präpositionen bei John Wycliffe und John Purvey*. Marburg: Diss. Universität Marburg.

Gasner, Ernst. 1891. *Beiträge zum Entwickelungsgang der neuenglischen Schriftsprache auf Grund der mittelenglischen Bibelversionen, wie sie auf Wyclif und Purvey zurückgehen sollen*. Nürnberg: Diss. Universität Göttingen.

Hoey, Michael. 1993. *Data, description, discourse: Papers on the English language in honour of John McH Sinclair*. London: HarperCollins.

Hopper, Paul J. & Elisabeth Closs Traugott. 2002. *Grammaticalization*, 2. Aufl. Cambridge: Cambridge University Press.

Hunston, Susan. 2002. *Corpora in applied linguistics*. Cambridge: Cambridge University Press.

Jespersen, Otto. 1949. *A modern English grammar on historical principles: Sounds and spellings*. Kopenhagen: Munksgaard.

Johansson, Stig. 1978. *Manual of information to accompany the Lancaster-Oslo Bergen Corpus of British English, for use with digital computers*. Oslo: Oslo University Press.

Kucera, Henry & W. Nelson Francis. 1967. *Computational analysis of present-day American English*. Providence: Brown University Press.

Lakoff, George. 2004. *Don't think of an elephant: Know your values and frame the debate*. White River Junction: Chelsea Green.

Leech, Geoffrey. 1998. Preface. In Sylvianne Granger, *Learner English on computer*, xiv-xx. London: Longman.

Leech, Geoffrey. 1992. Corpora and theories of linguistic performance. In Jan Svartvik & Randolph Quirk (Hrsg.), *Directions in corpus linguistics*, 105-122. Berlin: Mouton de Gruyter.

Mair, Christian & Marianne Hundt (Hrsg.). 2000. *Corpus linguistics and linguistic theory: Proceedings of ICAME 20*. Amsterdam: Rodopi.

Meyer, Charles F. 2002. *English corpus linguistics*. Cambridge: Cambridge University Press.

Mindt, Dieter. 1995. *An empirical grammar of the English verb: Modal verbs*. Berlin: Cornelsen

Neumann, Stella (Hrsg.). 2008. Das Projekt CroCo. http://fr46.uni-saarland.de/croco/(20.03.2008.)
Partington, Alan. 1998. *Patterns and meanings: Using corpora for English language research and teaching.* Amsterdam: Benjamins.
Poos, Deanna & Rita Simpson. 2002. Cross-disciplinary comparisons of hedging: Some findings from the Michigan Corpus of Academic Spoken English. In Randi Reppen, Susan M. Fitzmaurice & Douglas Biber (Hrsg.), *Using corpora to explore linguistic variation.* 3-23. Amsterdam: Benjamins.
Quirk, Randolph. 1962. *The use of English*. London: Longman.
Quirk, Randolph & Sidney Greenbaum. 1982. *A university grammar of English*. London: Longman.
Rühlemann, Christoph. 2007. *Conversation in context: A corpus-driven approach*. London: Continuum.
Reppen, Randi, Susan M. Fitzmaurice & Douglas Biber (Hrsg.). 2002. *Using corpora to explore linguistic variation*. Amsterdam: Benjamins.
Schwarz, Kathrin & Philine Wilke. 2008. *Comparison of speeches of Hillary Clinton and Barack Obama.* Workshop Corpus Linguistics, Universität des Saarlandes (unveröffentlicht).
Svartvik, Jan & Randolph Quirk (Hrsg.). 1992. *Directions in corpus linguistics*. Berlin: Mouton de Gruyter.
Semino, Elena & Michael H. Short. 2004. *Corpus stylistics: speech, writing and thought presentation in a corpus of English writing.* London: Routledge
Sinclair, John. 1991. *Corpus, concordance, collocation.* Oxford: Oxford University Press.
Sinclair, John. 1995. *Collins Cobuild English grammar.* London: HarperCollins.
Sinclair, John (Hrsg.). 2004. *How to use corpora in language teaching.* Amsterdam: Benjamins.
Stenström, Anna-Brita, Gisle Andersen & Ingrid Kristine Hasund. 2002. *Trends in teenage talk: Corpus compilation, analysis and findings.* Amsterdam: Benjamins.
Stubbs, Michael. 2001. *Words and phrases: Corpus studies of lexical semantics.* Oxford: Blackwell.
Svartvik, Jan (Hrsg.). 1990. *The London-Lund corpus of spoken English: Description and research.* (Lund studies in English 82). Lund: Lund University Press.
Teubert, Wolfgang & Anna Cermakova. 2007. *Corpus linguistics.* London: Continuum.
Tognini-Bonelli, Elena. 2001. *Corpus linguistics at work.* Amsterdam: Benjamins.
Tsui, Amy Bik May. 2004. What teachers have always wanted to know – and how corpora can help. In John Sinclair (Hrsg.), 2004. *How to use corpora in language teaching*, 39-61. Amsterdam: Benjamins.

Stefan Diemer
Fakultät 1, Institut für Sprache und Kommunikation
TU Berlin
Straße des 17. Juni 135
D - 10623 Berlin

s.diemer@umwelt-campus.de