Die Häufigkeitsverteilung von Artikeln über bestimmte Namen und Begriffe in der Wochenzeitung DIE ZEIT (Jahrgänge 1995-2001)

Franz-Josef Elmer
Franz-Josef.Elmer@unibas.ch

11. Mai 2002

1 Einführung

Die CD-ROM aller Artikel der Wochenzeitung DIE ZEIT aus sieben Jahrgängen eröffnet die einzigartige Möglichkeit, die Häufigkeit von Wörter in einer Zeitung zu untersuchen. Leider ist es nicht möglich, den vollen Text direkt zu untersuchen, da er verschlüsselt ist. Die Suchmaschine liefert nur eine Liste, der Artikel, die das in der Suchmaske eingegeben Wort enthalten. D.h. die im folgenden gezeigten Diagramme zeigen nur die Häufigkeit der *Zeitungsartikel*, die dieses Wort mindestens einmal benutzen.

Diese Häufigkeitsverteilungen spiegeln

- das Auf- und Ab- von Modewörtern sowie
- gesellschaftliche und politische Aktualitäten wieder.

Trotz des recht kurzen Zeitraums von sieben Jahren hat meine Untersuchung zu interessanten und zum Teil auch erstaunlichen Resultaten geführt. In diesem Artikel werden die Statistiken von 66 Begriffen und Namen präsentiert.

2 Modewörter

Interessant sind Wörter, deren Gebrauch in den letzten Jahren deutlich zu- oder abnahm. Dabei betrachten wir zunächst nur die Moden im Sprachgebrauch. Auf Änderungen von Worthäufigkeiten auf Grund gesellschaftlichen Änderungen oder politischen Ereignissen werden wir weiter unten eingehen (siehe die Abschnitte 5 und 6).

2.1 Modewörter von heute

Welche Wörter sind heute modern? Einige sind in Abbildung 1 gezeigt. Die Zahl der Artikel, die diese Wörter verwenden, hat sich im Laufe der sieben Jahre ungefähr verdoppelt. Für das Wort *cool* oder auch *Klamotten* ist dies nicht verwunderlich. Dass die aber auch für *blöd* und *komplett* zutrifft, ist doch erstaunlich.

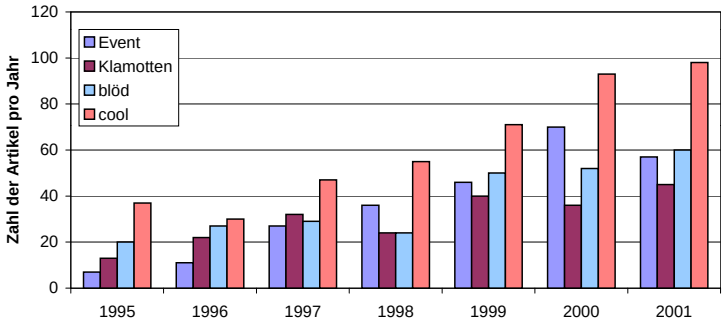

Abbildung 1: Die Artikelhäufigkeit für *blöd*, *cool*, *Klamotten* und *Event*.

2.2 Modewörter von gestern

Modewörter von heute oder gar von gestern zu finden, ist nicht ganz einfach (siehe Anhang A), insbesondere wenn sie keine konkreten gesellschaftlichen oder politischen Ereignisse oder Themen zugehören sollen[1].

Die Abbildungen 2 und 3 zeigen Beispiele, in denen die Häufigkeit der Artikel im Laufe der sieben Jahre um mehr als die Hälfte (bis auf *Zorn*) abgenommen hat. Der Niedergang des Wortes (bzw. Vorsilbe) *Öko* ist dabei nicht verwunderlich und ein Spiegelbild gesellschaftlicher Entwicklung.

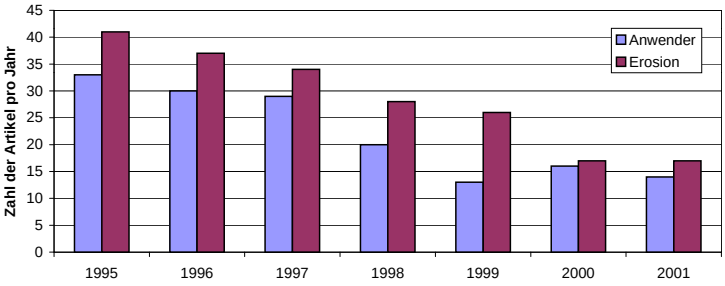

Abbildung 2: Die Artikelhäufigkeit für *Anwender* und *Erosion*.

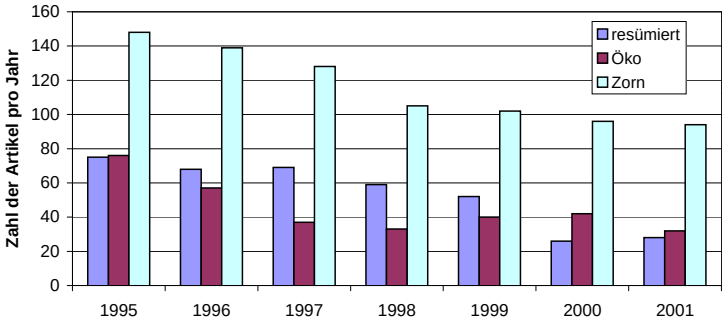

Abbildung 3: Die Artikelhäufigkeit für *resümiert*, *Öko* und *Zorn*.

Erinnern Sie sich noch an das Wort *Elchtest*? Es tauchte plötzlich im Herbst 1997 auf, brachte es zu einem Eintrag im DUDEN (allerdings nur als Terminus technicus der Autoindustrie) und — verschwand wieder (Abb. 4).

[1] Natürlich ist der Übergang fließend zumal Änderungen im Sprachgebrauch auch ein gesellschaftliches Phänomen sind.

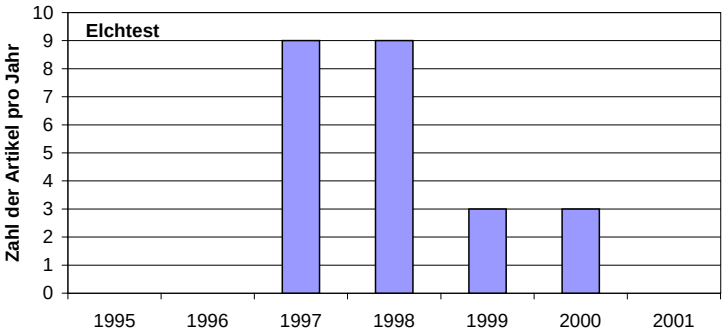

Abbildung 4: Die Artikelhäufigkeit für *Elchtest*.

3 Neue Rechtschreibung

In die Zeitspanne von 1995-2002 fällt die Einführung der neuen Rechtschreibung. Diese wurde im Sommer 1996 beschlossen und am 1.8.1998 offiziell eingeführt. In der ZEIT sind ab August 1999 alle Artikel gemäß der neuen Rechtschreibung verfasst. Die Abbildungen 5 und 6 dokumentieren diesen Wechsel anhand zweier Beispiele.

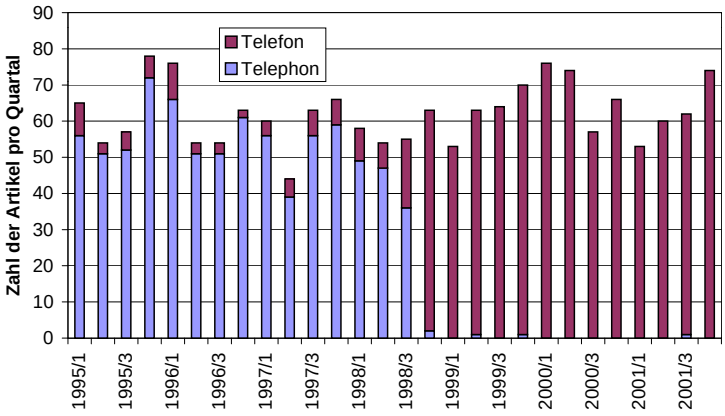

Abbildung 5: Die Artikelhäufigkeit für *Telephon* und *Telefon*.

Der Wechsel ist sehr deutlich, wobei der Übergang von *Telephon* zu *Telefon* ungefähr mit dem offiziellen Einführungstermin zusammenfällt, während der Wechsel von *Potential* zu *Potenzial* erst im Sommer 1999 erfolgte. Weiter ist bemerkenswert, dass *Telefon* schon vor 1998 hin und wieder benutzt wurde und die alte Form nach dem Sommer 1998 fast völlig ausgemerzt wurde. Dagegen gibt es vor dem zweiten Quartal 1999 keinen Artikel mit *Potenzial* dafür aber einige (über 10) mit der alten Form danach.

Ein anderes überraschendes Phänomen ist die Unterdrückung der Zahlworte nach 1998. Die Abbildungen 7 und 8 zeigen dies an den Beispielen *achtzig* versus *80* und *zwanzig* versus *20*. Weder im Regelwerk der neuen Rechtschreibung (Quelle: DUDEN) noch in den ZEITeigenen Regeln[2] habe ich einen Hinweis gefunden. Ist das eine ZEITinterne Regel um Platz zu sparen oder gibt es einen anderen Grund?

[2]Siehe den Artikel "Neue Rechtschreibung in der ZEIT" von Dieter E. Zimmer in der Ausgabe vom 10.6.1999

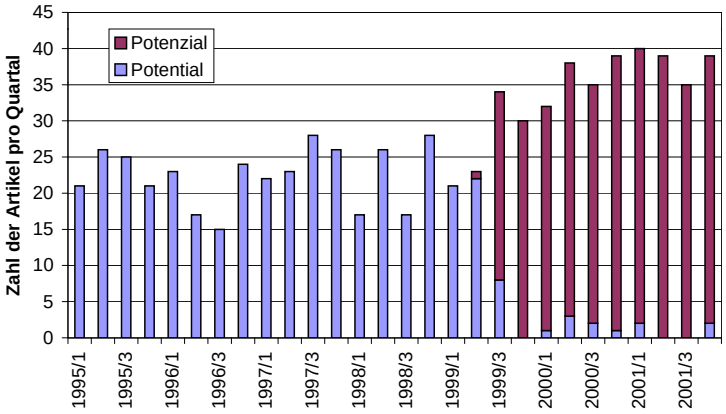

Abbildung 6: Die Artikelhäufigkeit für *Potential* und *Potenzial*.

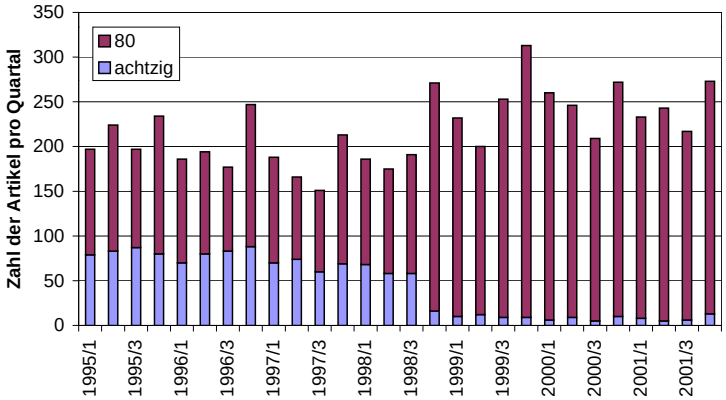

Abbildung 7: Die Artikelhäufigkeit für *achtzig* und *80*.

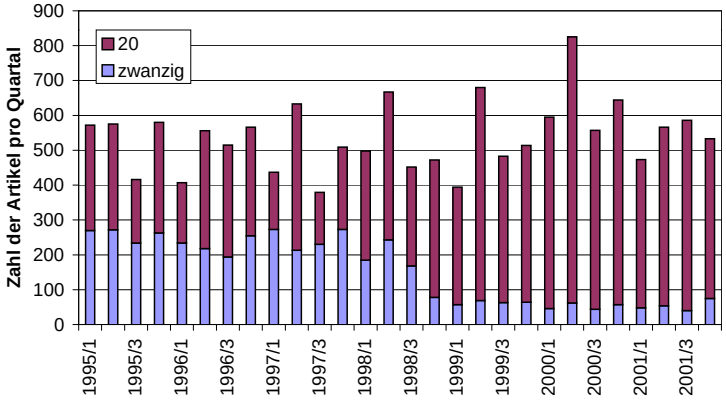

Abbildung 8: Die Artikelhäufigkeit für *zwanzig* und *20*.

4 Monatsnamen

Die Häufigkeitsverteilung der Monatsnamen bietet eine besondere Überraschung. Betrachten wir z.B. den *Oktober*: Abbildung 9 zeigt, dass jedes Jahr ungefähr gleich viele Artikel mit diesem Wort erscheinen.

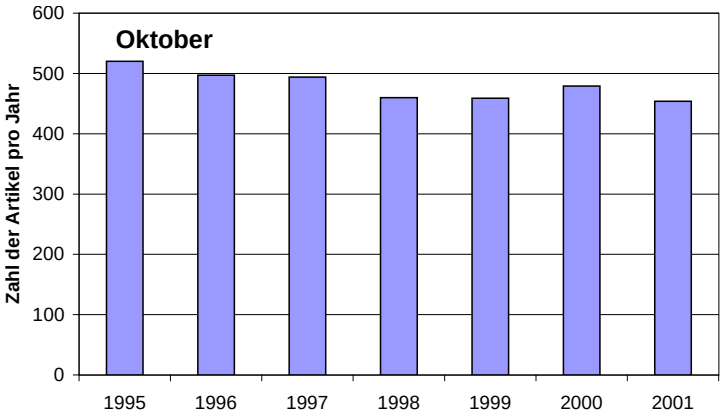

Abbildung 9: Die Artikelhäufigkeit für *Oktober* pro Jahr.

Bei einer Auflösung der Statistik auf der Monatsskala zeigt sich dagegen ein Auf und Ab der Häufigkeit (Abb. 10): Im Monat Oktober gibt es die meisten Artikel während ein halbes Jahr davor bzw. danach die wenigsten Artikel mit dem Wort *Oktober* erschienen. Auffallend ist auch die Symmetrie der Verteilung. In den beiden Monaten vor und nach dem Oktober ist die Häufigkeit überdurchschnittlich und in etwa gleich groß im September und November.

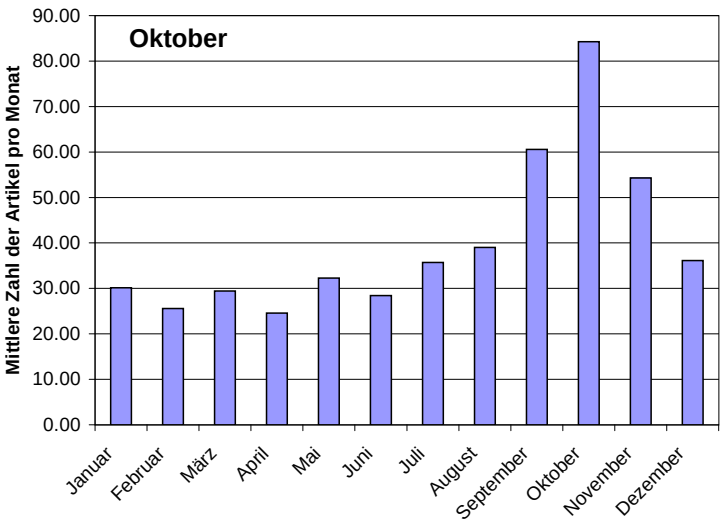

Abbildung 10: Mittlere monatliche Artikelhäufigkeit für *Oktober*.

Dieses Muster ist für alle Monatsnamen zu finden. Wegen den Terroranschlägen vom 11.9.2001 tanzt allerdings der *September* aus der Reihe (Abb. 11)

Analoge Schwankungen in der Häufigkeit findet man auch für *Weihnachten*, *Ostern* etc.

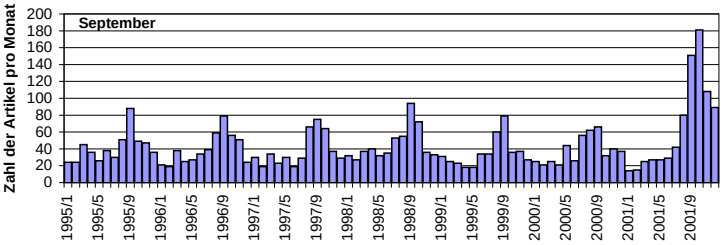

Abbildung 11: Die Artikelhäufigkeit für *September* pro Monat.

5 Gesellschaftliche Trends

In den letzten Jahren gab es starke gesellschaftliche Entwicklungen, die zum Teil durch rasante technologische Neuerungen möglich wurden. Zu diesen neuen Technolgien gehören gewisse Schlüsselwörter, deren Zunahme im Sprachgebrauch diese gesellschaftlichen Trends widerspiegelt.

Dies lässt sich auch in der Häufigkeitsentwicklung in der ZEIT sehr schön nachweisen. In diesem Kapitel werden die Häufigkeitsverteilung für einige dieser Schlüsselwörter aus den Bereichen

- Informationstechnologie (siehe Abschnitt 5.1)
- Telekommunikationstechnologie (siehe Abschnitt 5.2)
- Biotechnologie (siehe Abschnitt 5.3)

gezeigt.

Gerade die rasante Entwicklung dieser Hochtechnologien steht im Zusammenhang mit einer starken wirtschaftlichen Entwicklung. Im Abschnitt 5.4 werden einiger dieser Trends im Spiegel der Häufigkeitsverteilungen gezeigt.

Es gibt aber auch heiß diskutierte Themen, die wieder in der Versenkung verschwinden, wie wir an Beispielen im Abschnitt 5.5 sehen werden.

5.1 Computer und Internet

Die gesellschaftlich einflussreichste Hochtechnologie ist die Informationstechnologie vorallem durch das *Internet*. Abbildung 12 zeigt einen kontinuierlichen Anstieg der Zahl der Artikel in der ZEIT, die die Worte *Internet*, *Mail* oder *online* enthalten.

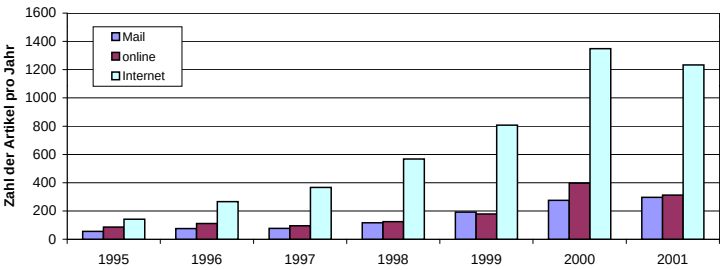

Abbildung 12: Die Artikelhäufigkeit für *Internet*, *online* und *Mail*.

Im Zusammenhang mit dem *Web* (einem wichtigen Aspekt des Internets) sind auch ganz neue Begriffe aufgetaucht, wie *website* und *homepage* (Abb. 13). Es fällt auf, dass trotz des starken Anstiegs der Häufigkeit dieser neuen Begriffe aus der Informationstechnologie, die Zahl der Artikel mit *Internet, online, Web* und *website* im Jahre 2001 geringer ist als im Jahr zuvor. Diese Abnahme zeigt sich auch für Begriffe aus der Wirtschaft (siehe Abschnitt 5.4). Der Grund dafür könnte das Platzen der Spekulationsblase "new economy" sein.

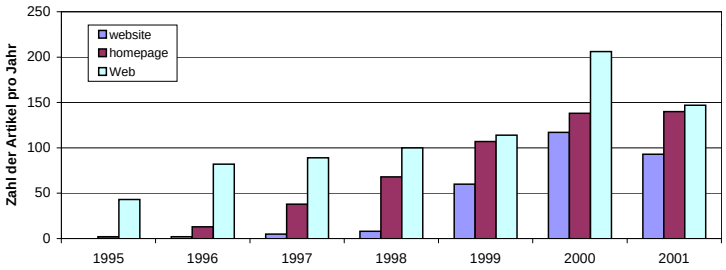

Abbildung 13: Die Artikelhäufigkeit für *Web*, *website* und *homepage*.

Andere Begriffe aus der Welt der Computer werden dagegen immer weniger gebraucht (Abb. 14). Z.B. ist das Wort *Computernetz* schon fast völlig durch *Internet* verdrängt worden. Und was eine *Diskette* ist, werden in zehn Jahren wahrscheinlich nur noch die Historiker der Informationstechnologie wissen.

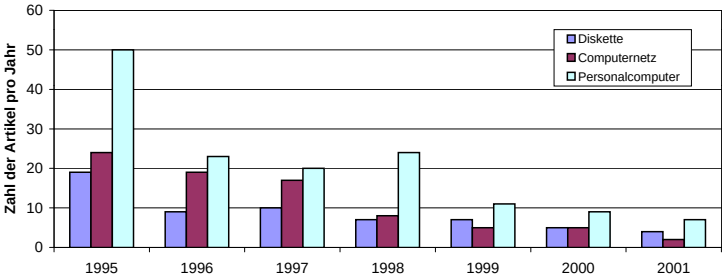

Abbildung 14: Die Artikelhäufigkeit für *Personalcomputer*, *Diskette* und *Computernetz*.

5.2 Mobilfunk

In der Telekommunikationstechnologie hat der *Mobilfunk* insbesondere in der Form seiner Endgeräten, den so genannten *Handys*, ungeahnte gesellschaftliche Entwicklungen ausgelöst. Natürlich führte das auch zu einem starken Anstieg von ZEIT Artikeln, in denen diese Begriffe verwendet werden.

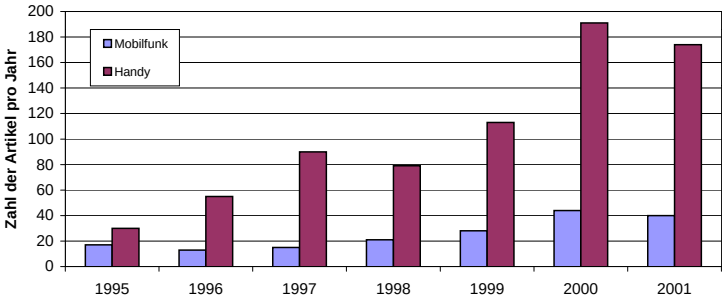

Abbildung 15: Die Artikelhäufigkeit für *Mobilfunk*, und *Handy*.

Eine wichtige neue Möglichkeit, die ein Handy bietet, ist das Verschicken von kurzen Textnachrichten via *SMS* (Short Message Service). Das Surfen im Web via *WAP* (Wireless App-

lication Protocol) ist wieder am Abklingen und wird sicher obsolet, wenn es die neuen *UMTS* (Universal Mobile Telecommunications System) Handys gibt. Alle diese technischen Kürzel haben in Jahr 2000 sehr stark zugenommen (Abb. 16). Artikel mit der Abkürzung *WAP* sind im Jahr 2001 aber auch wieder stark zurückgegangen[3]. Auch *SMS* scheint rückläufig zu sein. Das mag aber täuschen, denn im Jahr 2000 sind 11 mal "SMS ZEIT LEBEN Kurzleitartikel" von Roger de Weck erschienen. Entfernt man diese aus der Statistik, dann hat die Häufigkeit im Jahr 2001 im Vergleich zum Vorjahr zugenommen.

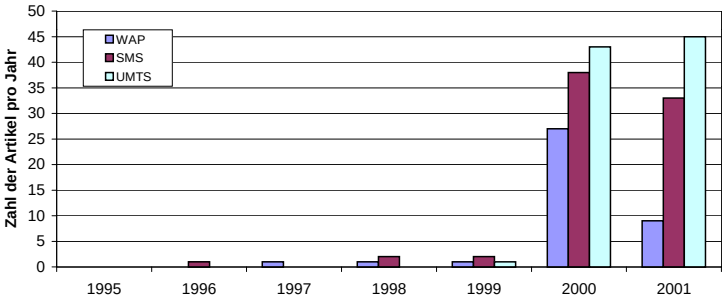

Abbildung 16: Die Artikelhäufigkeit für *WAP*, *SMS* und *UMTS*.

5.3 Biotechnologie

Zur kontroversesten aller Technologien ist die Biotechnologie geworden. Zwar gibt es diese Diskussionen um diese Technologie schon mehr als ein Jahrzehnt. Doch die Abbildung 17 zeigt, dass innerhalb der letzten sieben Jahren sich die Zahl der Artikel in der ZEIT, die diesen Begriff benutzen, verfünffacht hat. Diese Zunahme ist sogar noch stärker für Schlüsselwörter von besonders heiß umstrittenen Techniken, wie etwa *klonen* und *Stammzellen*. Das führte besonders im Jahr 2001 zu einer starken Zunahme von Artikeln, die den Begriff *Bioethik* enthalten.

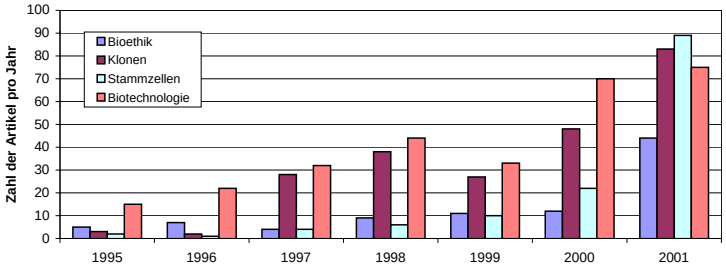

Abbildung 17: Die Artikelhäufigkeit für *Bioethik*, *klonen*, *Stammzellen*, und *Biotechnologie*.

5.4 Wirtschaft

Auch in der Wirtschaft hat es starke Änderung in der Häufigkeit bestimmter Begriffe gegeben. Wir betrachten drei Themengruppen.

Die erste Gruppe spiegelt die historische Entwicklung der europäischen Währung wider (Abb. 18): Der *Ecu*, die europäische Verrechnungseinheit, nahm an Bedeutung ab und so auch die Artikelhäufigkeit in der ZEIT. Im gleichen Maße nahm die Häufigkeit von *Euroland* zu, ein Begriff, der erstmals in der Ausgabe vom 17.10.1997 benutzt wurde.

[3] Übrigens musste für die in der in Abb. 16 gezeigten Häufigkeitsverteilung von *WAP* alle Artikel entfernt werden, die durch das Autorenkürzel "wap" gekennzeichnet waren.

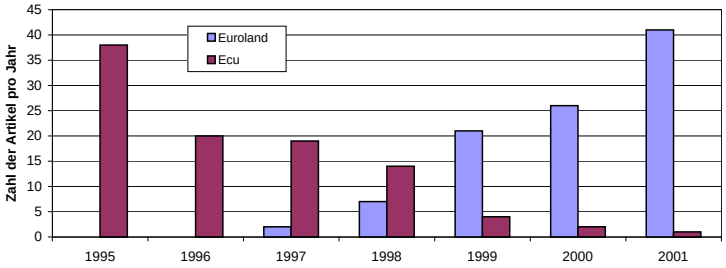

Abbildung 18: Die Artikelhäufigkeit für *Ecu* und *Euroland*.

Abbildung 19 zeigt vier Beispiele aus dem Themenbereich Aktiengesellschaften. Hier fällt auf, dass die Häufigkeit für alle Begriffe bis und mit 2000 zum Teil rasant anstieg aber im Jahre 2001 wieder abnahm. Besonders frappant ist diese Entwicklung für den Begriff *Shareholder value* verlaufen: Im Jahr 1995 gibt es nur einen Artikel mit diesem Begriff. Im Jahr 2000 sind es schon 60, aber im darauffolgenden Jahr weniger als die Hälfte.

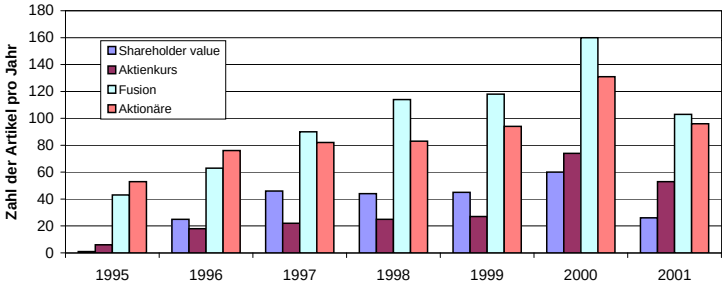

Abbildung 19: Die Artikelhäufigkeit für *Shareholder value*, *Aktienkurs*, *Fusion* und *Aktionäre*.

Der Rückgang in der Artikelhäufigkeit dieser Begriffe hängt vermutlich auch mit dem Zusammenbruch der Aktienkurse im Bereich der *new economy* zusammen. Dieser Begriff erscheint zum ersten Mal in der Ausgabe vom 14. August 1997. Abbildung 20 zeigt die Häufigkeiten dieses Begriffs sowie der damit eng verwandten Begriffen *e-commerce* (erstmals in der Ausgabe vom 30.7.1998 erwähnt) und *e-business* (erstmals am 4.6.1998 verwendet). Man beachte, dass in der Statistik auch solche Artikel mitgezählt werden, die zwar die einzelnen Worte eines zusammengesetzten Begriffs enthalten, aber nicht den Begriff selbst[4].

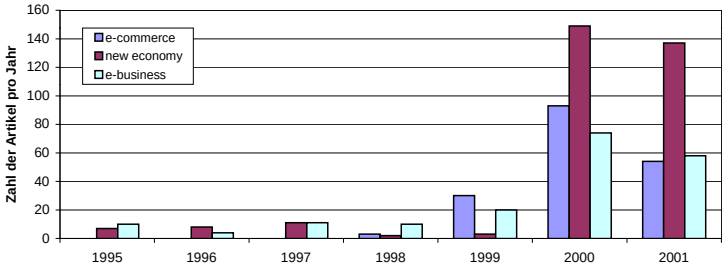

Abbildung 20: Die Artikelhäufigkeit für *e-commerce*, *new economy* und *e-business*.

[4]In den meisten Artikeln in denen es nicht um e-business und new economy geht, sind mit den Wörter *business* und *economy* Sitzplatzkategorien in Passagierflugzeugen gemeint.

5.5 Beispiele von abklingenden Aktualitäten

Es gibt auch Begriffe und Namen, welche plötzlich auftauchen, in aller Munde sind und dann wieder allmählich verschwinden. Abbildung 21 zeigt dafür drei Beispiele aus ganz unterschiedlichen Bereichen:

- Der Name *Goldhagen* tauchte das erste Mal in der Ausgabe vom 12. April 1996 auf und steht für die heftige Debatte, die Daniel Goldhagens Buch "Hitlers willige Vollstrecker" auslöste.
- *Viagra*, die Pille gegen Erektionsstörungen, wurde erstmals in der Ausgabe vom 15. Februar 1998 erwähnt.
- George Orwell prägte in seinem Roman "1984" den Begriff *Big Brother* für den alles überwachenden und beobachtenden Staat. Doch Ende 1999 (zuerst in Holland) und Anfang 2000 haben Fernsehsendungen in Deutschland gleichen Namens für Diskussionsstoff gesorgt. In dieser neuen Bedeutung wurde der Begriff erstmals in der Ausgabe vom 18. November 1999 verwendet. Alle älteren Artikel benutzen *Big Brother* im Orwellschen Sinne[5].

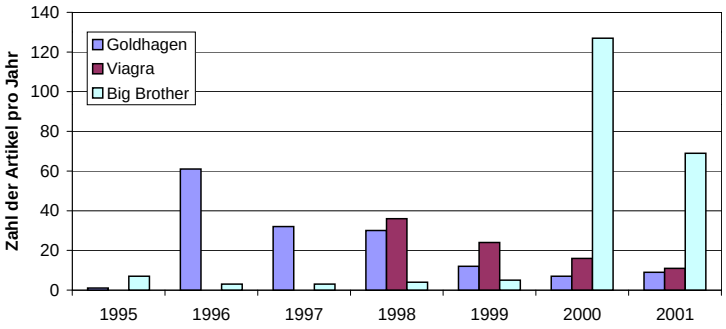

Abbildung 21: Die Artikelhäufigkeit für *Goldhagen*, *Viagra* und *Big Brother*.

6 Politische Aktualitäten

Häufigkeitsverteilungen von Artikeln bezüglich Personen und Schlüsselwörtern aus der Politik spiegeln sehr gut die politischen Aktualitäten wider. Es können im Folgenden nur einige wenige Beispiele aus der deutschen, europäischen sowie internationalen Politik präsentiert werden.

6.1 Deutsche Politik

Abbildung 22 zeigt die Häufigkeit der Artikel, die sich mit dem jetzigen und dem letzten Bundeskanzler beschäftigen. Beide erreichen im Wahljahr 1998 sehr hohe Werte. In der Regel wird eine Person in einer wichtigen politischen Rolle häufiger erwähnt im Vergleich zu der Zeit vor oder nachdem sie diese Rolle eingenommen hat. Doch im Jahr 2000 führt die CDU-Spendenaffäre für *Kohl* zu einer Ausnahme dieser Regel.

[5] In der Häufigkeitsverteilung gibt es auch eine verschwindend geringe Zahl von Artikel, in denen sowohl die Wörter *Big* und *Brother* vorkommen, aber nicht der Begriff *Big Brother*.

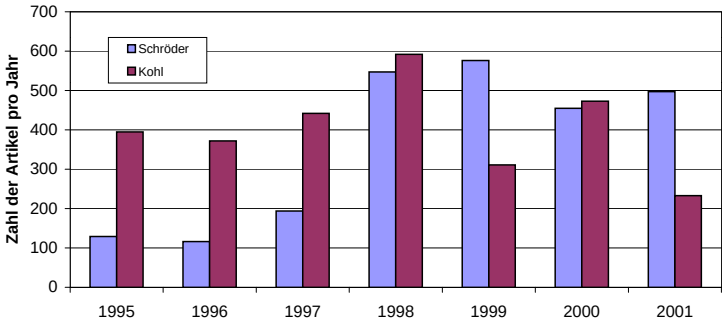

Abbildung 22: Die Artikelhäufigkeit für *Kohl* und *Schröder*.

Der Umzug der bundespolitischen Bühne von der ehemaligen Haupstadt *Bonn* nach *Berlin* zeigt sich klar in der Abnahme bzw. Zunahme der Artikelhäufigkeiten (Abb. 23). Dabei ist die Summe in etwa konstant geblieben.

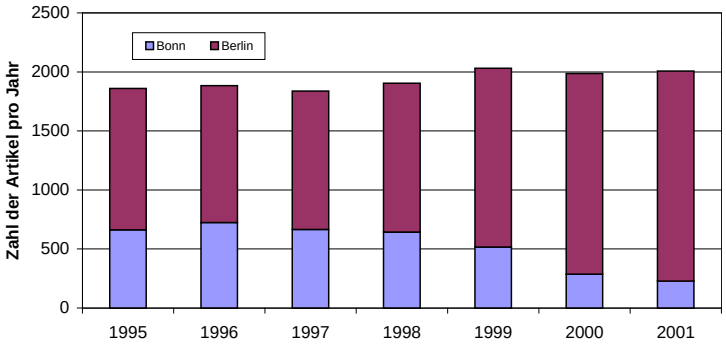

Abbildung 23: Die Artikelhäufigkeit für *Bonn* und *Berlin*.

6.2 Europäische Politik

Wissen Sie noch wann der Kosovokonflikt hochkochte? Ein Blick auf die Abbildung. 24 gibt die Antwort: Es war im Frühjahr 1999. Das Maximum ist im April, zu der Zeit als die NATO ihre Angriffe gegen Serbien begann. Die Häufigkeit der Artikel mit dem Begriff *NATO* ist genau in der Zeit des Krieges besonders hoch. Auch die Frage, ob die NATO *Bodentruppen* einsetzen soll oder nicht, spiegelt sich in einer Häufigkeitsverteilung wider, welche im Frühjahr 1999 in etwa parallel zur Häufigkeit von *Kosovo* verläuft.

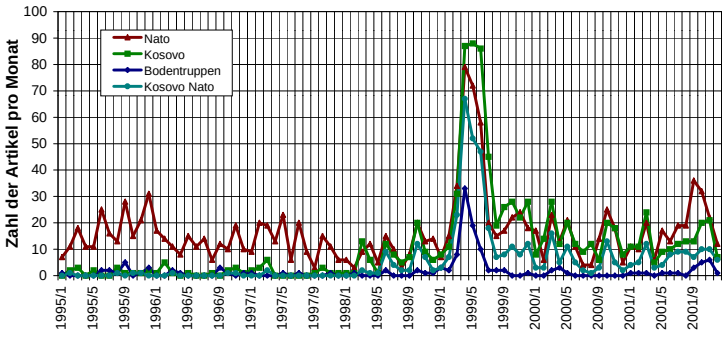

Abbildung 24: Die Artikelhäufigkeit für *Kosovo*, *NATO*, *Bodentruppen* sowie die Kombination von *Kosovo* und *NATO*.

Auffällig ist die starke Korrelation zwischen den Häufigkeitsverteilungen von *Kosovo* und *NATO* ab dem Frühjahr 1998. Abbildung. 25 zeigt, dass die Häufigkeitverteilung der Artikel,

die zwar *NATO* aber nicht *Kosovo* enthalten, für 1999 keine besonderen Merkmale zeigt. Dagegen hat umgekehrt die Häufigkeitsverteilung für Artikel, die *Kosovo* aber nicht *NATO* enthalten ein klares Maximum im Frühjahr 1999.

Ein weiteres erstaunliches Merkmal zeigt die *NATO* Verteilung vor dem April 1999: Es gibt oszillatorische Schwankungen in der Artikelhäufigkeit mit einer geschätzten Periode von etwa drei Monaten. Diese Oszillation ist besonders regelmäßig im Jahr 1995. Was steckt da wohl dahinter?

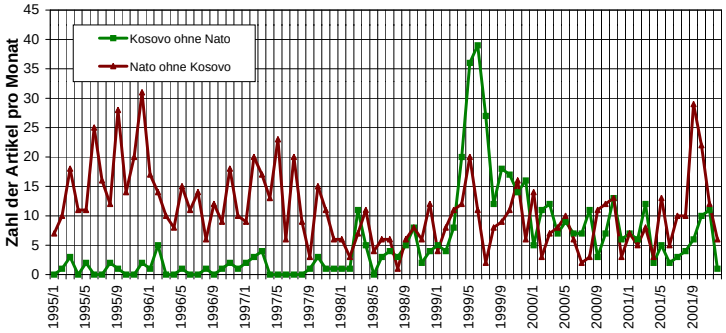

Abbildung 25: Die Zahl der Artikel mit *Kosovo* aber ohne *NATO* und umgekehrt.

Abbildung 26 versucht den Wechsel im Amt des russischen Präsidenten von *Jelzin* zu *Putin* in der Artikelstatistik widerzugeben. Putin tauchte im August 1999 aus dem politischen Nichts auf, als er vom amtierenden Präsidenten Jelzin zum Ministerpräsidenten ernannt wurde. Zum Jahreswechsel 1999/2000 trat Jelzin sein Amt an Putin ab. Betrachtet man die Häufigkeitsverteilung zu *Jelzin* alleine, so erkennt man dieses plötzliche Abtreten Jelzins von der politischen Bühne nicht. Vielleicht spiegelt die allmähliche Abnahme der Artikel, die sich mit *Jelzin* beschäftigen, die immer geringer werdende Bedeutung, die er gegen Ende seiner Präsidentschaft spielte, wider.

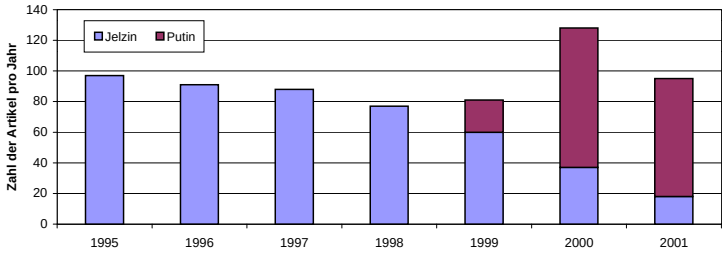

Abbildung 26: Die Artikelhäufigkeit für *Jelzin* und *Putin*.

6.3 Außereuropäische und Internationale Politik

Die Abbildung 27 zeigt Beispiele, die sich um Clinton drehen. Da ist zum eine das Auf und Ab der Häufigkeitsverteilung zum Namen *Clinton* selber. Interessant dabei ist, dass es 1997 nur wenig mehr Artikel gab als 2001, das Jahr in dem Clinton offiziell nur die ersten 20 Tage bis zur Amtsübergabe an seinen Nachfolger regierte. Die meisten Artikel sind im Jahr 1998 erschienen. Der Grund dafür hat mit der Clinton-Lewinsky Affäre zu tun: Monica Lewinsky war eine Praktikantin im Weissen Haus, die behauptete, eine sexuelle Affäre mit Clinton gehabt zu haben, was dieser leugnete. Der Name *Lewinsky* tauchte in der Ausgabe vom 29. Januar 1998 zum ersten Mal auf. Die Affäre nahm solche Ausmaße an, dass ein Absetzungverfahren (engl. impeachment) eingeleitet wurde. Der erste Artikel, welcher den Begriff *Impeachment* in diesem Zusammenhang erwähnte, erschien am 6. August 1998.

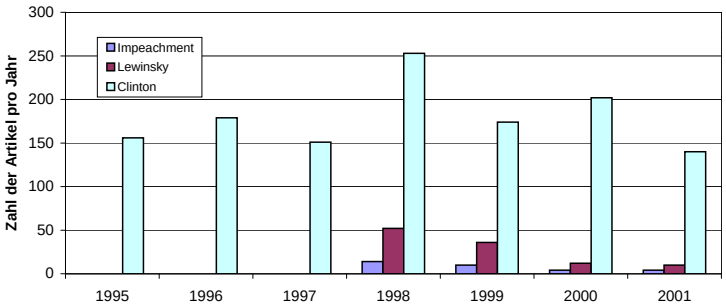

Abbildung 27: Die Artikelhäufigkeit für *Clinton*, *Lewinsky* und *Impeachment*.

Die Häufigkeitsverteilungen von *Lewinsky* und *Impeachment* zeigen einen charakteristischen Verlauf für kurzfristige gesellschaftliche oder politische Modethemen (siehe auch Abb. 21): Ein steiler Anstieg (innerhalb weniger Wochen) und ein allmählicher Zerfall mit einer "Halbwertszeit" von etwa einem Jahr.

Der Konfliktherd Naher Osten wird in Abbildung 28 durch die Statistiken für *Israel, Arafat* und *Intifada* repräsentiert. Man erkennt, dass die Häufigkeitsentwicklung aller drei Wörter in etwa parallel verlief: Bis zum Jahr 1999 nahm die Häufigkeit leicht (für *Israel*) bis stark (für *Arafat* und *Intifada*) ab. Seither haben alle in starken Maße zugenommen. Dies hängt mit dem Ausbruch der so genannten zweiten Intifada im Herbst 1999 zusammen.

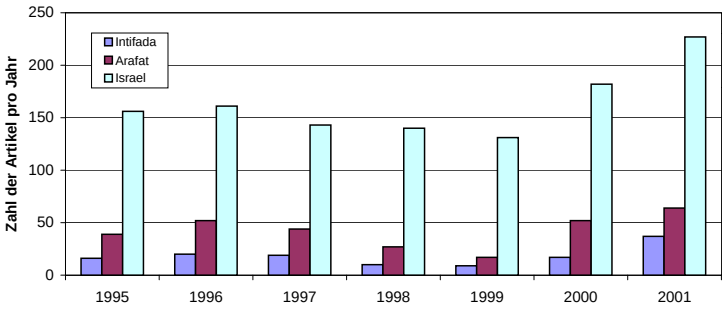

Abbildung 28: Die Artikelhäufigkeit für *Israel*, *Arafat* und *Intifada*.

Die Abbildungen 29 und 30 spiegeln die die Wahl von *Bush* zum Präsidenten der USA, die *Terror*-Anschläge vom 11. September 2001 sowie deren Konsequenzen für das *Taliban*-Regime in Afghanistan wider.

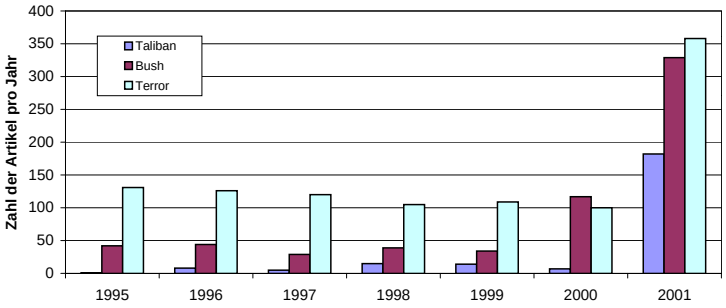

Abbildung 29: Die Artikelhäufigkeit für *Terror*, *Bush* und *Taliban*.

Alle drei Wörter erfahren eine Vervielfachung der Artikelhäufigkeit nach dem 11. September, welche aber den typischen Abfall danach zeigt.

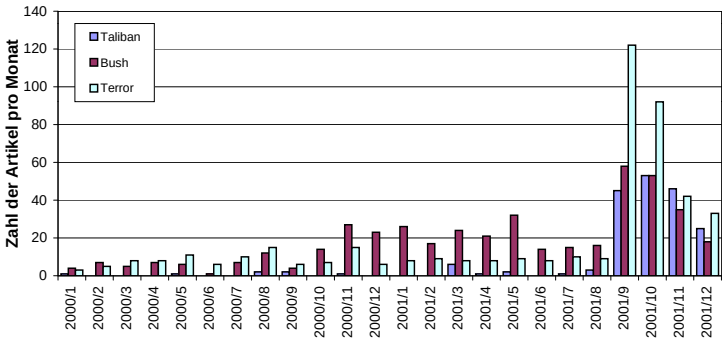

Abbildung 30: Die Artikelhäufigkeit für *Terror*, *Bush* und *Taliban* in den Jahren 2000 und 2001.

A Methodik

Seitdem ich DIE ZEIT auch als CD-ROM abonniert habe, haben mich die Worthäufigkeiten schon immer interessiert. Leider war es ohne Kenntnis der benutzten Software nicht möglich, an den Text oder an die Indexdatei zu gelangen. Ich habe dieses Projekt deshalb zunächst auf die lange Bank geschoben.

Die Situation hat sich mit den beiden letzten Ausgaben der CD-ROM geändert. Jetzt gibt es ein Programm auf der CD-ROM, welches die Suchmaschine als lokalen Web Server betreibt (URL-Adresse: `http://localhost:3075/start`). Im Web Browser steht eine Eingabemaske für den Suchbegriff zur Verfügung. Nach dessen Eingabe liefert der Web Server eine Tabelle mit den Suchergebnissen. Ich habe nun ein kleines Program (`ZEITSearch`) in Java geschrieben, welches nach Eingabe eines Suchbegriffes solche eine Suche initiiert und das Suchergebnis statistisch auswertet und grafisch (als Balkendiagramm) darstellt. Die Statistik lässt sich in ein bekanntes textuelles Format (CSV) abspeichern, welches von allen gängigen Tabellenkalkulationsprogrammen weiterverarbeitet werden kann.

Die Suchmaschine auf der ZEIT CD-ROM liefert alle Artikel, die den Suchbegriff enthalten, egal wo er im Artikel vorkommt. Jeder Artikel beginnt mit dem Datum, den Wörtern “DIE ZEIT” und dem Titel der Rubrik (falls vorhanden). Danach erst folgt der Titel des Artikels und der Artikeltext. Am Ende gibt es eine Liste von Schlüsselwörter. Diese Worte müssen im Artikel selber gar nicht vorkommen. Man muss deshalb bei der Interpretation der Statistiken vorsichtig sein. Besonders abrupte Änderungen in der Artikelhäufigkeit lassen darauf schließen, dass ein Schlüsselwort oder eine Rubrik verschwand oder neu eingeführt wurde. Die Abbildung 31 zeigt diese Problematik am Bespiel des Wortes *kompakt*. Die Häufigkeit der Artikel, die dieses Wort mindestens einmal enthalten nimmt stark ab. Eine genauere Analyse des Suchergebnisses der Suchmaschine zeigt allerdings, dass es bis zum Mai 1999 ein Rubrik mit dem Titel “CD kompakt” gab. Nimmt man aus der Statistik alle Artikel, die sowohl *CD* als auch *kompakt* enthalten, so ergibt sich sogar eine Zunahme der Artikel mit *kompakt*.

Zusätzlich zu `ZEITSearch` habe ich noch ein weiteres Java Programm geschrieben (`ZEITKeyWordStatistics`) welches

1. automatisch für alle Wörter einer Wortliste ein Suchanfragen auf der Suchmaschine startet,
2. die Antwort statistisch auswertet und
3. die Zahl der Artikel pro Jahr in eine Datenbank abspeichert.

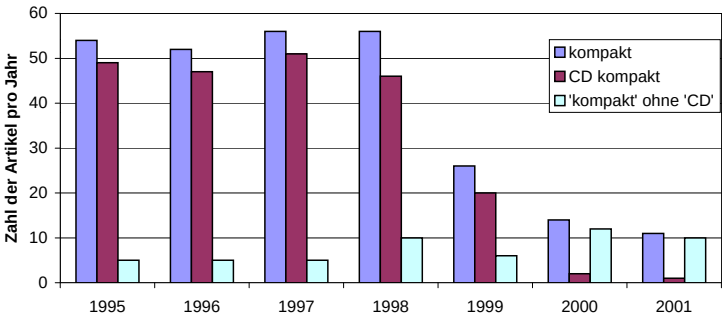

Abbildung 31: Die Artikelhäufigkeit für *kompakt*.

Eine geeignete Wortlist habe ich auf der ZEIT CD-ROM mit den Jahrgängen 1995-1999 gefunden. Diese Liste enthält zu jedem Wort auch die Zahl der Artikel in diesen fünf Jahrgängen. Aus dieser List von ca. 700 000 Wörtern (sowie Abkürzungen und Zahlen) habe ich nur die Wörter ausgewählt, die in mindestens 10 Artikel vorkamen. Mit dieser abgespeckten List von etwa 90 000 Wörter lief dann das Programm `ZEITKeyWordStatistics`. Mit der gezielten Suche nach z.B. ansteigenden oder absteigenden Häufigskeitsverteilungen habe ich dann nach interessanten Wörter gesucht. Insbesondere die Wörter, deren Statistiken in den Abbildungen 1-3, 7 und 8 gezeigt sind, habe ich auf diese Weise gefunden.