Das Internet als linguistisches Korpus

Hans Bickel (Basel)

Abstract

This article discusses whether the Internet can be used as a linguistic corpus. It is based on experiences in connection with the *Variantenwörterbuch des Deutschen* (Dictionary of Standard German Variants), which was compiled 1997-2004. In order to identify national and regional variants of the German language in Germany, Austria and Switzerland, it was necessary to work with a large linguistic corpus that could also provide data on the frequency of rather rare words. The question was: Is the Internet suitable as a corpus for linguistic frequency analysis? The use of the WWW as corpus can be suitable only

1. if reliable and reproducible results can be obtained;
2. if the results are closely related to the language as it is actually used.

The test showed that the Internet is an extremely useful corpus to get information on word frequency. The enormous size and the large number of different text types makes it an extremely versatile corpus, which has a systematic connection to the written language reality.

1 Einleitung

Seit leistungsstarke Computer mit grosser Speicherkapazität zu einem normalen Arbeitsinstrument geworden sind, ist die Korpuslinguistik zu einem populären Zweig der Sprachwissenschaft geworden (vgl. z.B. Lemnitzer 2006). Voraussetzung für korpuslinguistische Fragestellungen ist ein möglichst grosses Textkorpus. Als Prototyp eines solchen Korpus gilt gewöhnlich das zwischen 1991 und 1994 entstandene British National Corpus (BNC)[1] mit 100 Mio. Textwörtern. Seither sind in grösserer Zahl weitere Korpus-Projekte entstanden oder am entstehen (z. B. Czech National Corpus,[2] Birmingham Bank of English,[3] International Corpus uf English,[4] Penn Treebank Project[5]).

Für das Deutsche sind im Wesentlichen die folgenden Korpora zu nennen:

- das grosse Cosmas-Korpus am Institut für deutsche Sprache IdS in Mannheim,[6] das fast zwei Milliarden Textwörter enthält;

[1] http://info.ox.ac.uk/bnc.

[2] http://ucnk.ff.cuni.cz/english/index.html.

[3] http://www.titania.bham.ac.uk.

[4] http://www.ucl.ac.uk/english-usage/ice/.

[5] http://www.cis.upenn.edu/~treebank/home.html.

[6] http://www.ids-mannheim.de/kt/corpora.shtml.

- das Gutenbergprojekt, das literarische Texte von über 1000 Autoren, deren Werke nicht mehr urheberrechtsgeschützt sind, in gemeinsamer Anstrengung der Internet-Gemeinschaft sammelt;[7]
- das Projekt *Digitales Wörterbuch der deutschen Sprache (DWDS)*[8] in Berlin mit den Partnerprojekten *Schweizer Text Korpus*[9] in Basel und *Austrain Academy Corpus (AAC)*[10] in Wien.

Neben diesen systematisch aufgebauten Korpora gibt es mit dem Internet und den über das WWW-Protokoll abrufbaren Internetseiten ein riesiges, organisch gewachsenes Archiv mit mehreren Milliarden Textseiten, das dank den entsprechenden Suchmaschinen auch als Korpus verstanden werden kann. Das Internet ist letztlich ein täglich wachsendes, sich verändernde Korpus, das von Millionen von Internetnutzern aufgrund ganz unterschiedlicher Bedürfnisse und Interessen zusammengestellt wird. Und da der Aufbau eines systematischen Sprachkorpus ausgesprochen aufwändig und entsprechend teuer ist, stellt sich die Frage, ob nicht auch das Internet als bereits bestehende, in der Grösse alle anderen Korpora übertreffende Sammlung von Texten als Korpus für korpuslinguistische Fragestellungen genutzt werden kann.

Ausgangspunkt meiner Überlegungen sind die Erfahrungen bei der Arbeit am *Variantenwörterbuch des Deutschen* (vgl. Ammon et al. 2006). Für dieses Wörterbuch wurden rund 12'000 Wörter und Wendungen gesammelt, deren Gebrauch entweder national oder regional begrenzt oder unterschiedlich ist, sowie die gemeindeutschen, also im gesamten Sprachraum gebräuchlichen Entsprechungen dieser Varianten.

2 Die Korpusgrundlage des Variantenwörterbuchs

Eine ganz entscheidende Voraussetzung für die Erstellung eines Wörterbuches ist ein Quellenkorpus, das als Grundlage für die Belegsammlung ausgewertet wird und auf dem die Angaben im Wörterbuch beruhen. Das Korpus ist natürlich abhängig von der Sprachwirklichkeit, die abgebildet werden soll. Wenn das Ziel wie in diesem Fall lautet, alle standardsprachlichen nationalen Varianten mit Ausnahme der eigentlichen Fachsprache zu sammeln, sollte das Korpus eine repräsentative Auswahl standardsprachlichen Schreibens enthalten. Nun ist aber die deutsche Sprachwirklichkeit so gross und komplex, dass es keine formalen Verfahren gibt, um eine statistisch repräsentative Auswahl daraus zu treffen. Eine Quellenauswahl kann daher immer nur eine bessere oder schlechtere Annäherung an eine tatsächlich repräsentative Auswahl sein. Wie gut diese Annäherung ist, darüber kann mangels Überblickbarkeit der Grundgesamtheit nur spekuliert werden.

Für die Erstellung des Variantenwörterbuchs galt es also, in Texten aus den deutschsprachigen Ländern nach Wörtern und Wendungen mit national oder regional begrenzter Verwen-

[7] http://gutenberg.spiegel.de.

[8] http://www.dwds.de.

[9] http://www.schweizer-textkorpus.ch.

[10] http://www.aac.ac.at/.

dung zu suchen. Dazu wurde ein traditionelles, nicht-digitales Quellenkorpus erstellt. Wir haben uns entschieden, um möglichst die gesamte Breite des schriftlichen Standardsprachgebrauchs auszuwerten, dieses Quellkorpus anhand einer Liste von sachlichen Domänen zusammenzustellen und zu jeder dieser Domänen eine Anzahl Quellen auszuwerten. Damit sollten die wichtigsten Bereiche standardlichen Schreibens abgedeckt werden. Die Domänen reichten von der *Körperpflege* über *Handwerk, Jugendkultur, Kochkunst* bis hin zu *Post, Politik, Verwaltung*, usw. Dazu kamen noch aus jedem Land mindestens 50 neuere Romane, 10 Kriminalromane, 10 Trivialromane, 50 Tages- und Wochenzeitungen und ca. 50 Zeitschriften, Illustrierte, Magazine.[11]

In diesen Quellen wurden alle nicht gemeindeutschen Wörter der Standardsprache markiert und in eine Datenbank aufgenommen. Die Erfahrung zeigt, dass mit diesem "traditionellen" Verfahren viele der bekannten und auch einige neue nationale Varianten bereits nach verhältnismässig kurzer Zeit belegt werden können. Wesentlich schwieriger wird es jedoch, wenn auch quantitative Aussagen gemacht werden sollen. Denn an ein Korpus der nationalen Varianten werden andere Anforderungen gestellt als an ein konventionelles Wörterbuchkorpus.

Im Gegensatz zu einem "normalen" Wörterbuch, in dem alle gebräuchlichen Wörter verzeichnet werden und daher nur entschieden werden muss, ob ein Wort lexikalisiert ist oder ob es sich lediglich um eine Augenblicksbildung handelt, bot die Arbeit am Variantenwörterbuch einige zusätzliche Schwierigkeiten. Aufgenommen wurden ja nicht einfach die Wörter, die in einem bestimmten Zentrum vorkommen. Vielmehr musste gesichert sein, dass ein Wort zusätzlich in einem anderen Zentrum oder in grösseren Teilen des eigenen Zentrums *nicht* vorkommt. Es reichte also nicht festzustellen, dass beispielsweise das Wort *Bostitch* ein in der deutschen Schweiz gebräuchliches Wort ist, sondern es galt gleichzeitig herauszufinden, dass dieses Wort in Deutschland oder Österreich nicht gebraucht wird, dass in Deutschland *Tacker* und *Hefter*, in Österreich dazu *Klammerlmaschine* gesagt wird. Nötig war also nicht nur ein positiver Test, der die Existenz eines Wortes in einem bestimmten Gebiet feststellte, sondern ebenso ein negativer Test, der das Vorkommen des Wortes an anderen Orten ausschloss.

Diese doppelte Bedingung stellte ganz besondere Anforderungen an das methodische Vorgehen bei der Belegsammlung und an den Umfang und die Ausgewogenheit des Korpus. Die Wörterbuchmitarbeiterinnen und -mitarbeiter konnten sich nur zur Hälfte auf ihre Sprachkompetenz verlassen, nämlich bei der Feststellung, ob ein Wort in ihrem Zentrum vorkam. Zur Beurteilung der nationalen oder geographischen Reichweite waren sie auf Hilfe von aussen angewiesen.

Aus diesem Grund wurden in einem Bearbeitungszentrum ausschliesslich die Quellen der jeweils beiden anderen Zentren gelesen. Dabei wurden alle sprachlichen Erscheinungen angestrichen und kommentiert, die im jeweils eigenen Zentrum nicht vorkamen. Nur auf diese Weise bestand Gewähr, dass der zweite, negative Test zuverlässig durchgeführt werden konnte.

[11] Eine genaue Übersicht findet sich in Ammon et al. (2006: 911 ff.).

Dieses an sich bewährte Verfahren konnte aber dennoch nicht absolute Sicherheit bezüglich der richtigen Beurteilung von Varianten bieten. Schwierigkeiten ergaben sich einerseits daraus, dass der Wortschatz jedes Menschen beschränkt ist. Besonders in Sachbereichen, die jemanden nicht besonders interessieren, bestehen meist auch deutliche Lücken im Wortschatz. Wer sich nicht für Wirtschaft interessiert, dem sind Ausdrücke wie *Cashflow* und *Abschreibung* nicht geläufig. Andererseits gibt es, und dies trifft in besonderer Weise auf Deutschland zu, auch Wörter, die nur in Teilbereichen eines Landes gebräuchlich sind. Einem Norddeutschen ist es aber nicht unbedingt bewusst, dass *Sonnabend* nur in einem Teilgebiet Deutschlands gesagt wird. Es ist daher für einen Einzelnen nahezu unmöglich, bei einem bestimmten Wort dessen Verbreitungsgebiet zu bestimmen.

Dazu kam, und dies galt insbesondere für die Schweizer Bearbeiterinnen und Bearbeiter, dass die Standardsprache der anderen Zentren durch Literatur und Medien meist doch einigermassen häufig rezipiert wird, so dass viele Wörter der anderen Zentren, die man aktiv in der Schweiz nicht brauchen würde, doch vielen sehr bekannt vorkamen. Das Schweizer Team war manchmal nicht einmal ganz sicher, ob es ein deutschländisches oder österreichisches Wort nicht vielleicht selbst auch brauchen würden.

Um diese Unsicherheiten auszugleichen, war es notwendig, mit einem verlässlichen Korpus zu arbeiten, aus dem einigermassen gesicherte Angaben über das Vorkommen eines Lexems gezogen werden konnten. Nun ist aber der Aufbau eines systematischen Korpus, das auch Angaben über Wortfrequenzen liefert, eine äusserst zeitraubende und in Zeiten knapper Forschungsmittel manchmal fast unmögliche Angelegenheit. Zwar erhält man bereits bei einigen tausend Belegen Auskunft über die häufigsten nationalen Varianten. Zur Abschätzung von Frequenz und kleinräumigem Gültigkeitsbereich braucht es aber Belegdatenbanken mit mehreren Millionen Textwörtern. Eine solches Korpus, das unseren Bedürfnissen genügt hätte, gab und gibt es für das Deutsche noch nicht und es war im vorgesehenen und finanzierbaren Rahmen ausgeschlossen, ein solches zu erstellen.

Wir haben daher bereits Ende der 90er Jahre begonnen, das Internet oder genauer das World Wide Web als Textkorpus zu nutzen, indem wir die Suchmaschinen, zuerst AltaVista, später auch Google, als Korpusabfragesystem verwendeten.

3 Internet-Suchmaschinen

Im Juni 1999 waren ungefähr 330 Millionen Internetseiten bei AltaVista indiziert. Damit stand bereits damals ein ausserordentlich umfangreiches Korpus zur Verfügung, das ohne weiteren Aufwand vom Forscher genutzt werden konnte. Allerdings war natürlich nur der kleinere Teil der Internetseiten auf Deutsch verfasst. Vorherrschende Sprache war und ist das Englische.

Suchmaschinen erlauben, nach Seiten in einer gewünschten Sprache zu suchen. Und nicht nur das, man kann bei den meisten Suchmaschinen auch die so genannte *Internetdomain* bestimmen, in der gesucht werden soll. Die Adressen im WWW sind in verschiedene *Domains* gegliedert. Diese werden im letzten Teil einer Internetadresse ausgedrückt. Die Adresse der Universität Basel lautet beispielsweise *http://www.unibas.ch*. Die Universität gehört damit zur

Domain *ch,* eine Domain, zu der die meisten schweizerischen Anbieter von Internetseiten gehören.[12] Deutschländische Seiten finden man unter der Domain *de,* österreichische unter *at*.

Dank dieser Einteilung in Domains ist es möglich, im Internet gezielt nach der Frequenz von einzelnen Wörtern in unterschiedlichen Ländern zu suchen. Insgesamt standen zu Beginn unserer Forschungen bei AltaVista ungefähr 20 Millionen deutsche Seiten zur Verfügung. Dabei muss man bedenken, dass eine Seite nur ganz wenige Wörter enthalten kann, während andere mehrere A4-Textseiten umfassen. Seither hat sich die Zahl der Seiten stark erhöht. Eine Abfrage Ende Juni 2006 ergab die folgende Anzahl deutscher Seiten für die deutschsprachigen Länder:

Anzahl bei Suchmaschinen indizierter deutschsprachiger Internetseiten nach Länderdomains (30.6.2006)		
Domain	AltaVista	Google
Domain .at	55.5 Mio	87.1 Mio
Domain .ch	53.2 Mio	115 Mio
Domain .de	724 Mio	934 Mio
Domain .be (für Ostbelgien)	6.35 Mio	1.99 Mio
Domain .li (für Liechtenstein)	0.627 Mio	2.29 Mio
Domain .lu (für Luxemburg)	0.965 Mio	1.95 Mio
Domain .it (für Südtirol)	11.7 Mio	4.4 Mio
Total	**852 Mio**	**1'147 Mio**

Tabelle1: Die Tabelle zeigt die Anzahl indizierter deutschsprachiger WWW-Seiten bei den Suchmaschinen AltaVista und Google, aufgeteilt für die nationalen Voll- und Halbzentren des Deutschen.

4 Die deutschen WWW-Seiten als lexikographisches Korpus

Die Hauptfrage war aber: Eignet sich das Internet als Basis für lexikographische und sprachstatistische Untersuchungen? Gewöhnlich wird für die lexikographische Forschung ein nach sorgfältigen Kriterien aufgebautes, in sich konsistentes Korpus benutzt. Im WWW dagegen ist ein Korpus von Sprachdaten vorhanden, das von niemandem in seiner Gesamtheit überblickt werden kann. Dazu war es bereits zu Beginn unserer Untersuchungen mit seinen über 20 Millionen deutschen Seiten schlicht zu umfangreich und mit seinen täglich neu dazukommenden Seiten zu dynamisch.

Sicher konnte ein derart diffuses und unüberblickbares Korpus nicht einfach bedenkenlos ausgewertet werden. Die unglaubliche Grösse und der geringe Aufwand, den es zu seiner Auswertung brauchte, waren aber zu verlockend, um ohne weiteres darauf zu verzichten. Die Nutzung des WWW als Quelle konnte jedoch nur dann in Frage kommen, wenn

[12] Ausnahmen waren in den neunziger Jahren fast nur die grossen multinationalen Konzerne mit Sitz in der Schweiz oder auch die internationalen gemeinnützigen Institutionen, wie z. B. das Internationale Komitee vom Roten Kreuz. Heute sind natürlich weitere Domains dazugekommen, unter anderen *eu* oder *tv,* wodurch länderspezifische Abfragen deutlich erschwert werden.

1. erwiesen werden konnte, dass damit einigermassen zuverlässige und vor allem reproduzierbare Ergebnisse erzielt werden konnten;

2. wenn die Ergebnisse in einem systematischen Bezug zur Sprachwirklichkeit standen.

Um diese beiden Bedingungen zu überprüfen, haben wir Tests entwickelt, die auf dem bisherigen lexikographischen Wissen basierten. Sie sollten demonstrieren, wie stark dem Internet-Korpus vertraut werden durfte und wo allenfalls Abweichungen und Verzerrungen erwartet werden mussten.

Eine Schwierigkeit ergab sich insofern, dass es für das Deutsche kaum sprachstatistische Arbeiten gab, die für einen Vergleich herangezogen werden konnten. Der bisher einzige grössere Versuch von H. Meier (1967 [1964]) war nach über dreissig Jahren nicht mehr aktuell. Ein Vergleich war daher nur bedingt möglich.

Um Bedingung 1) zu überprüfen, wurden zehn Lemmata willkürlich ausgewählt, die in der Lexikographie bisher nicht als in irgendeiner Weise national geprägt angesehen wurden. Die Überprüfung dieser Wörter im Korpus von AltaVista musste also vergleichbare prozentuale Ergebnisse liefern. Dazu wurde diese Abfrage nach einiger Zeit wiederholt, um allfällige Veränderungen bei der ständigen Neuindexierung durch die Suchmaschine zu verfolgen. Die Ergebnisse sind in den Tabellen 2-4 dargestellt.

AltaVista Abfrageergebnisse vom 22.10.1998				
Lexem	A	CH	D	Gesamt
selten	4'691	5'643	42'465	52'799
	8.88%	10.69%	80.43%	100%
wollen	68'700	67'490	541'690	677'880
	10.13%	9.96%	79.91%	100%
Tisch	2'930	3'420	25'030	31'380
	9.34%	10.90%	79.76%	100%
Mensch	8'064	8'357	62'130	78'551
	10.27%	10.64%	79.10%	100%
Baum	1'843	1'580	16'322	19'745
	9.33%	8.00%	82.66%	100%
Kopf	6'691	8'101	65'792	80'584
	8.30%	10.05%	81.64%	100%
soll	81'040	64'010	624'390	769'440
	10.53%	8.32%	81.15%	100%
schön*	20'106	21'662	168'835	210'603
	9.55%	10.29%	80.17%	100%
Regen	1'392	1'929	14'517	17'838
	7.80%	10.81%	81.38%	100%
Computer	83'050	111'320	813'770	1'008'140
	8.24%	11.04%	80.72%	100%
Total Abs.	**278'507**	**293'512**	**2'374'941**	**2'946'960**
Total %	**9.45%**	**9.96%**	**80.59%**	**100%**

AltaVista Abfrageergebnisse vom 25.2.1999				
Lexem	A	CH	D	Gesamt
selten	4'592	5'360	42'575	52'527
	8.74%	10.20%	81.05%	100%
wollen	67'410	63'440	532'470	663'320
	10.16%	9.56%	80.27%	100%
Tisch	2'850	3'338	25'305	31'493
	9.05%	10.60%	80.35%	100%
Mensch	8'152	8'213	62'063	78'428
	10.39%	10.47%	79.13%	100%
Baum	1'998	1'693	17'271	20'962
	9.53%	8.08%	82.39%	100%
Kopf	6'636	8'063	67'776	82'475
	8.05%	9.78%	82.18%	100%
soll	76'470	59'990	616'800	753'260
	10.15%	7.96%	81.88%	100%
schön*	19'494	20'141	166'394	206'029
	9.46%	9.78%	80.76%	100%
Regen	1'056	1'388	10'716	13'160
	8.02%	10.55%	81.43%	100%
Computer	80'160	103'240	763'242	946'642
	8.47%	10.91%	80.63%	100%
Total Abs.	**268'818**	**274'866**	**2'304'612**	**2'848'296**
Total %	**9.44%**	**9.65%**	**80.91%**	**100%**

AltaVista Abfrageergebnisse vom 25.2.2006				
Lexem	A	CH	D	Gesamt
selten	1'140'000	912'000	18'600'000	20'652'000
	5.52%	4.42%	90.06%	100%
wollen	5'560'000	4'150'000	83'500'000	93'210'000
	5.97%	4.45%	89.58%	100%
Tisch	2'250'000	1'740'000	33'300'000	37'290'000
	6.03%	4.67%	89.30%	100%
Mensch	3'210'000	2'310'000	51'300'000	56'820'000
	5.65%	4.07%	90.29%	100%
Baum	1'860'000	1'280'000	32'700'000	35'840'000
	5.19%	3.57%	91.24%	100%
Kopf	3'180'000	2'500'000	53'100'000	58'780'000
	5.41%	4.25%	90.34%	100%
soll	8'060'000	6'180'000	113'000'000	127'240'000
	6.33%	4.86%	88.81%	100%
schön	3'170'000	2'480'000	53'200'000	58'850'000
	5.39%	4.21%	90.40%	100%
Regen	1'650'000	1'290'000	39'500'000	42'440'000
	3.89%	3.04%	93.07%	100%

Computer	6'310'000 3.56%	5'160'000 2.91%	166'000'000 93.54%	177'470'000 100%
Total Abs.	**36'390'000**	**28'002'000**	**644'200'000**	**708'592'000**
Total %	**5.14%**	**3.95%**	**90.91%**	**100%**

Tabelle 2-4: In obigen Tabellen sind die Anzahl Seiten und die jeweiligen Prozentangaben aufgeführt, die die Abfrage zu drei verschiedenen Zeitpunkten bei der Internet-Suchmaschine AltaVista für die drei nationalen Zentren Österreich (A), Schweiz (CH) und Deutschland (D) ergeben hat. In Österreich und der Schweiz befanden sich Ende der neunziger Jahre je ca. 10% der deutschsprachigen Internetseiten, in Deutschland je ungefähr 80%. 7 Jahre später haben sich die prozentualen Verhältnisse verschoben. die österreichischen Seiten sind mit ca. 5% der Gesamtmenge der deutschen Seite vertreten, Schweizer Seiten mit 4% und deutschländische Seiten mit ca. 90%. Mit Asterisk bezeichnete Lemmata (Bsp. *schön*) werden bei der Suche auch in ihren Flexionsformen gefunden.

Die Ergebnisse bei diesen zehn ausgewählten Wörtern zeigen deutlich, dass bei national nicht markierten Wörtern durchaus vergleichbare Resultate zustande kommen. Die prozentualen Werte lagen Ende der 90er Jahre des 20. Jahrhunderts für Österreich und die Schweiz im Schnitt bei ungefähr 9.5%, für Deutschland bei 80.5%. Allerdings haben sich bis heute, d.h. sieben Jahre später, die Verhältniszahlen doch um einiges verschoben. Gemeindeutsche Wörter kommen in Österreich nur noch auf eine durchschnittliche Frequenz von 5%, in der Schweiz sogar nur noch auf 4%, während sich die Zahlen in Deutschland auf durchschnittlich 91 Prozent erhöht haben.

Die Abfragen zu verschiedenen Zeitpunkten machen dennoch zwei Dinge deutlich:

1. Das Internet als dynamisches Korpus verändert sich, über längere Zeiträume können sich die Prozentanteile verschiedener Domains verschieben. Angesichts der Tatsache, dass sich das Internet immer noch in einer dynamischen Wachstumsphase befindet ist das wenig erstaunlich.
2. Abfragen zu einem bestimmten Zeitpunkt sind jedoch in sich konsistent. Wörter ohne nationale Markierung erscheinen zu einem bestimmten Zeitpunkt in vergleichbaren Prozentzahlen.

Nachdem die Tests in den 90er Jahren des letzten Jahrhunderts gezeigt hatten, dass bei mehreren unterschiedlichen Lemmata immer wieder vergleichbare Resultate erzielt wurden, stellte sich die Frage, wie nationale Varianten im Internet-Korpus aufscheinen. Dazu wurden vier Lemmata ausgewählt, die nun aber in der Lexikographie eindeutig als national markiert beschrieben wurden. Es waren dies: *Maturand* (nach Duden 1996 'schweiz., sonst veraltet'), *Maturant* (nach Duden 1996 'österr.'), *Abiturient* (in Duden 1996 nicht markiert, jedoch bei Meyer 1989 unter dem Lemma *Maturand* als in der Schweiz ganz unüblich markiert), *allfällig* (nach Duden 1996 'bes. österr., schweiz.'). Die Resultate sind in den Tabellen 5 und 6 dargestellt.

ISSN 1615-3014

AltaVista Abfrageergebnisse vom 22.10.1998				
Lexem	A	CH	D	Gesamt
Maturand*	0	282	4	286
	0.00%	98.60%	1.40%	100 %
Maturant*	823	4	16	843
	97.63%	0.47%	1.90%	100 %
Abiturient*	26	31	3'953	4'010
	0.65%	0.77%	98.58%	100 %
allfällig*	2'369	6'335	317	9'093
	26.26%	70.23%	3.51%	100 %

AltaVista Abfrageergebnisse vom 25.2.2006				
Lexem	A	CH	D	Gesamt
Maturand	891	82'300	19'200	102'391
	0.87%	80.38%	18.75%	100%
Maturant	18'100	55	780	18'935
	95.59%	0.29%	4.12%	100%
Abiturient	4'290	161	378'000	382'451
	1.12%	0.04%	98.84%	100%
allfällig	16'100	78'900	1'520	96'520
	16.68%	81.74%	1.57%	100%

Tabellen 5 und 6: Wie in Tabellen 2-4 sind auch hier die Anzahl Seiten und die Prozentangaben aufgeführt, die die Abfrage bei der Internet-Suchmaschine AltaVista für die drei nationalen Zentren Österreich (A), Schweiz (CH) und Deutschland (D) ergeben hat. Im Unterschied zu Tabellen 2-4 sind allerdings nur Lemmata abgefragt worden, die in der Lexikographie bereits als national markiert gelten. Die Ergebnisse bestätigen die Angaben in den Wörterbüchern: Alle Lemmata zeigen deutlich nationale Verbreitungsschwerpunkte. *Abiturient* beispielsweise ist sowohl 1998 wie auch noch 2006 mit über 98% der Fundstellen vorwiegend in Deutschland verbreitet, *Maturand* mit zuerst ähnlicher prozentualer Häufung der Belege in der Schweiz, 2006 ist allerdings der prozentuale Anteil der Schweizer Seiten nur noch bei 80%, was aber immer noch 20 mal häufiger ist als bei gemeindeutschen Wörtern

Die Resultate in den Tabelle 5 und 6 sind eindeutig. Die Erwartungen aus dem lexikographischen Vorwissen wurden bestätigt und teilweise präzisiert. *Maturand, Maturant* und *Abiturient* waren tatsächlich fast ausschliesslich in jeweils einem nationalen Zentrum auf Internetseiten zu finden. Die wenigen Belege, die in den jeweils anderen Zentren gefunden wurden, stammten meist von WWW-Seiten, die Informationen über das Ursprungszentrum enthielten oder von Autoren aus diesem Zentrum geschrieben wurden. Da Österreicher, Schweizer und Deutsche in allen drei Zentren anzutreffen sind und in diesen zum Teil auch publizieren, gibt es selten 100-Prozent-Resultate. Einzelne 'Ausreisser' fanden sich, entsprechend der Sprachwirklichkeit, in allen Zentren. Sie liessen sich mit AltaVista auch einzeln anschauen und überprüfen, so dass sie in der Regel erklärt werden konnten.

Im Unterschied zu den Angaben in Duden 1996, wo das Wort *allfällig* als 'bes. österr., schweiz.' markiert ist, zeigte die AltaVista-Abfrage in aller Deutlichkeit, dass *allfällig* vor

allem in der deutschen Schweiz gebräuchlich ist, in Österreich schon deutlich seltener und fast gar nicht in Deutschland.

Sicher dürfen diese Ergebnisse nicht als bis auf die zweite Kommastelle mit der Sprachwirklichkeit im Einklang stehend interpretiert werden. Zu wenig fassbar sind sowohl die Gesamtmenge der Internetseiten als auch die Sprachwirklichkeit. Und die neuesten Abfragen zeigen, dass sich über die Jahre durchaus Frequenzverschiebungen von einigen Prozent ergeben. Die von uns durchgeführten Tests haben aber eindeutig ergeben, dass die erzielten Resultate einerseits konsistent und reproduzierbar sind und dass sie andererseits die aus der Lexikographie gewonnenen Ergebnisse bestätigen. Die Resultate dürfen daher guten Gewissens als Hinweise auf Frequenz und Vorkommen eines Wortes genommen werden. Das heisst nicht, dass man den Ergebnissen blind vertrauen darf. Als Ergänzung eines auf systematische Weise zusammengestellten Korpus liefern sie jedoch wesentliche Zusatzinformationen zu Verbreitung und Vorkommen der Wörter.

5 Ausblick bzw. Perspektiven einer Internet-Lexikographie-Forschung

Rückblickend hat sich gezeigt, dass das Internet ein äusserst brauchbares Korpus zur Frequenzabklärung darstellt. Die enorme Grösse und die Vielzahl unterschiedlicher Textsorten, sie reichen von persönlichen Homepages über Verwaltungs- und Gesetzestexte, wissenschaftliche Abhandlungen, Werbung und Dienstleistungsangebote, bis hin zu Zeitungs- und Zeitschriftenarchiven, machen es zu einem äusserst vielseitigen Korpus, das in einem systematischen Bezug zur verschriftlichten Sprachwirklichkeit steht.

Wie der Bezug genau ist, lässt sich allerdings nicht sagen. Dies liegt daran, dass es keine aktuelle Sprachstatistik des Gegenwartsdeutschen gibt und dass die systematischen Korpora noch zu klein sind, um verlässliche Frequenzangaben zu liefern. Versuche, die Liste der häufigsten deutschen Wörter nach Meier (1967: 112f.) mit den Internetresultaten zu vergleichen, haben teilweise Übereinstimmungen, zum Teil aber auch eklatante Unterschiede ergeben. So sind etwa die Wörter *ich* und *Paragraph* im Internet gegenüber Meiers Statistik massiv untervertreten, *Zeit, Menschen, Frau* dagegen weisen eine vergleichbare Frequenz auf.

Diese Unterschiede müssen nicht zwangsläufig auf eine Verzerrung des Internet-Korpus zurückgeführt werden. Eine gewisse Verzerrung könnte auch das Korpus von Meier 1967 aufweisen. Dazu sind seit seiner Untersuchung fast 40 Jahre vergangen, in denen wahrscheinlich auch Frequenzverschiebungen stattgefunden haben. Das Wort *Paragraph* wird wohl kaum noch zu den häufigsten Wörtern der Gegenwartssprache gehören.

Wichtigstes Ergebnis der Versuche ist jedoch, dass das Internet-Korpus in sich konsistent ist. Die Resultate sind nicht zufällig, sondern bilden die Sprachwirklichkeit auf dem Internet ab. Und diese Sprachwirklichkeit hat trotz der Flüchtigkeit des Mediums einen erstaunlich konstanten Charakter. Trotz der rasanten Vervielfachung der Internetseiten in den letzten Jahren hat sich an den erzielten Resultaten wenig geändert. Die Grösse und Vielfalt des Korpus garantierte seine Stabilität auch während der enormen Wachstumsphase.

ISSN 1615-3014

Da zu Beginn unserer Forschung die Suchmaschine von Google noch nicht existiert hat, wurden die Abfragen mit AltaVista gemacht. Tests zeigen allerdings, dass man mit Google heute zu ganz ähnlichen Ergebnissen kommt. Die prozentuale Verteilung für die gemeindeutschen Wörter ist heute bei Google noch vergleichbar mit den AltaVista-Angaben vom Ende der 90er Jahre. D.h. dass das Google-Korpus heute wohl in sich konistenter ist als das AltaVista-Korpus.

Google Abfrageergebnisse vom 7.7.2006				
Lexem	A	CH	D	Gesamt
selten	1'980'000	1'720'000	13'300'000	17'000'000
	11.65%	10.12%	78.24%	100%
wollen	4'690'000	4'390'000	55'600'000	64'680'000
	7.25%	6.79%	85.96%	100%
Tisch	1'560'000	1'710'000	10'200'000	13'470'000
	11.58%	12.69%	75.72%	100%
Mensch	2'430'000	2'420'000	20'000'000	24'850'000
	9.78%	9.74%	80.48%	100%
Baum	892'000	773'000	6'260'000	7'925'000
	11.26%	9.75%	78.99%	100%
Kopf	2'130'000	2'000'000	16'700'000	20'830'000
	10.23%	9.60%	80.17%	100%
soll	6'220'000	6'130'000	61'900'000	74'250'000
	8.38%	8.26%	83.37%	100%
schön	2'390'000	2'410'000	25'900'000	30'700'000
	7.79%	7.85%	84.36%	100%
Regen	1'030'000	862'000	6'570'000	8'462'000
	12.17%	10.19%	77.64%	100%
Computer	9'220'000	6'520'000	128'000'000	143'740'000
	6.41%	4.54%	89.05%	100%
Total Abs.	**32'542'000**	**28'935'000**	**344'430'000**	**405'907'000**
Total %	**8.02%**	**7.13%**	**84.85%**	**100%**

Tabelle 7: Eine aktuelle Abfrage mit der Suchmaschine Google zeigt, dass das Google-Korpus bez. der prozentualen geografischen Verteilung mit dem damals um Faktoren kleineren AltaVista-Korpus vom Ende der 90er Jahre weitgehend übereinstimmt.

Die konsistenten Ergebnisse haben den Ausschlag gegeben, einzelne Ausgaben von Tageszeitungen, Protokolle von Parlamentssitzungen und Romane zu umfangreichen Wortlisten zu verarbeiten und diese systematisch auf nationale Varianten mittels automatisierter Internet-Abfrage zu überprüfen. Dadurch konnten für das Variantenwörterbuch viele neue nationale oder regionale Varianten eruiert und bekannte Varianten empirisch abgesichert werden.

Die Möglichkeiten, die das Internet für die Lexikographie bietet, werden wohl in Zukunft von den meisten Wörterbuchprojekten der Standardsprache genutzt werden. Die leichte Zugänglichkeit eines fast unbeschränkt grossen Korpus, seine elektronische Form und die Automatisierungsmöglichkeiten für die Belegerfassung machen das WWW zu einer idealen Quelle für

die Wortschatzerforschung. Einzelne andere Wörterbuchprojekte haben bereits auch begonnen, das Internet systematisch einzubeziehen, und wollen auch Frequenzangaben zu den Lemmata liefern.[13] Es sind aber auch andere Forschungen denkbar, so etwa Wortschatzanalysen einzelner Autoren oder Textsorten im Hinblick auf die Verwendung von zentralem oder peripherem Wortschatz und im Hinblick auf stärkere oder schwächere nationale Markierung. Damit ist das Internet in kurzer Zeit nicht nur zu einem wichtigen Informationsmedium avanciert, sondern auch zu einer neuartigen, äusserst brauchbaren Quelle der linguistischen Forschung.

Literaturangaben

Ammon, Ulrich (1995): *Die deutsche Sprache in Deutschland, Österreich und der Schweiz: das Problem der nationalen Varietäten*. Berlin.

Ammon, Ulrich (1997). "Vorüberlegungen zu einem Wörterbuch der nationalen Varianten der deutschen Sprache". In: Moelleken, Wolfgang W.;/Weber, Peter J. (eds.): *Neue Forschungsarbeiten zur Kontaktlinguistik*. Bonn: 1-9.

Ammon, Ulrich/Bickel, Hans/Ebner, Jakob/Esterhammer, Ruth/Gasser, Markus/Hofer, Lorenz/Kellermeier-Rehbein, Birte/Löffler, Heinrich/Mangott, Doris/Moser, Hans/ Schläpfer, Robert†/Schlossmacher, Michael/Schmidlin, Regula/Vallaster, Günter (2006): *Variantenwörterbuch des Deutschen. Die Standardsprache in Österreich, der Schweiz und Deutschland sowie in Liechtenstein, Luxemburg, Ostbelgien und Südtirol*. Berlin.

Bickel, Hans (2000): "Das Internet als Quelle für die Variationslinguistik". In: Häcki Buhofer, Annelies (ed.): *Vom Umgang mit sprachlicher Variation. Soziolinguistik, Dialektologie, Methoden und Wissenschaftsgeschichte*. Festschrift zum 60. Geburtstag von Heinrich Löffler. Tübingen: 111-124.

Bickel, Hans/Schmidlin, Regula (2004): "Ein Wörterbuch der nationalen und regionalen Varianten der deutschen Standardsprache". In: Studer, Thomas/Schneider, Günther (eds.): *Deutsch als Fremdsprache und Deutsch als Zweitsprache in der Schweiz*. Neuchâtel: 99-122. (= *Bulletin suisse de linguistique appliquée* 79).

Bergmann, Rolf (ed.) (1988): *Probleme der Textauswahl für einen elektronischen Thesaurus. Beiträge zum ersten Göttinger Arbeitsgespräch zur historischen deutschen Wortforschung*. Leipzig.

Clyne, Michael (ed.) (1992): *Pluricentric Languages: Differing Norms in Different Nations*. Berlin/New York.

Duden (1996). *Deutsches Universalwörterbuch [A-Z]*. 3., neu bearb. und erw. Aufl. Mannheim/Zürich [etc.].

Ebner, Jakob (1998): *Wie sagt man in Österreich? Wörterbuch der österreichischen Besonderheiten*. 3., vollst. überarb. Aufl. Mannheim/Wien/Zürich. (= *Duden-Taschenbücher* 8).

Hofer, Lorenz (1999): "Ein Wörterbuch mit nationalen Varianten des Deutschen". *Sprachspiegel* 1: 7-15.

Lemnitzer, Lothar (2006): *Korpuslinguistik: eine Einführung*. Tübingen.

[13] Siehe z. B. die Zusammenstellung der Akademie der Wissenschaften zu Göttingen: http://grimm.adw-goettingen.gwdg.de/wbuecher/.

Meier, Helmut (1967): *Deutsche Sprachstatistik*. 2., erw. und verb. Aufl. Hildesheim.

Meyer, Kurt (1989): *Wie sagt man in der Schweiz? Wörterbuch der schweizerischen Besonderheiten*. Mannheim/Zürich. (= *Duden-Taschenbücher* 22).

Ris, Roland (1988): "Der schweizerische Anteil in den deutschen Grosswörterbüchern". In: Bergmann, Rolf (ed.): *Probleme der Textauswahl für einen elektronischen Thesaurus. Beiträge zum ersten Göttinger Arbeitsgespräch zur historischen deutschen Wortforschung*. Leipzig: 113-126.

ISSN 1615-3014