

Hans Bickel (Basel)

Das Internet als Quelle für die Variationslinguistik

Gliederung

1.	Einleitung	1
2.	Das Wörterbuchprojekt „Nationale Varianten der deutschen Standardsprache“	1
3.	Überlegungen zum Korpus und zur Korpusauswertung eines Wörterbuchprojektes	2
4.	Die Suchmaschine AltaVista	4
5.	Die deutschen WWW-Seiten als lexikographisches Korpus	5
6.	Ausblick bzw. Perspektiven einer Internet-Lexikographie-Forschung	8
7.	Anmerkungen	9
8.	Literatur	9

1. Einleitung

Die Frage, der in den folgenden Darlegungen nachgegangen werden soll, ist, welche neuen Möglichkeiten sich der Variationslinguistik dank der Entwicklung im Bereich des Internets eröffnen und welche neuen Quellen mit dem Internet erschlossen werden können. Ausgangspunkt der Überlegungen ist das gegenwärtig laufende Forschungsprojekt „Wörterbuch *Nationale Varianten der deutschen Standardsprache*“, in dessen Rahmen das Internet als Quelle zurzeit erprobt wird. Ziel dieses Projektes ist es, ein Wörterbuch zur erstellen, in dem möglichst alle nationalen und regionalen standardsprachlichen Varianten des Deutschen erfasst und dokumentiert werden. Im Folgenden wird kurz die methodische Voraussetzung dieses Projektes geschildert, um dann anhand der Überlegungen zum Korpus die Bedeutung des Internets als Ergänzung der herkömmlichen Belegsammlung zu diskutieren und erste Ergebnisse unserer Recherchen vorzustellen.

2. Das Wörterbuchprojekt „Nationale Varianten der deutschen Standardsprache“

Im Gegensatz zu älteren Auffassungen geht man heute in der Regel nicht mehr von einer einheitlichen Standardsprache aus, die in Deutschland richtig gesprochen und geschrieben wird und an ihren Rändern, d. h. vor allem in der Schweiz und Österreich zunehmend unreiner und variantenreicher wird. Vielmehr hat sich, auch im Anschluss an die angelsächsische Forschung (z. B. Clyne 1992), allmählich die Auffassung durchgesetzt, dass Sprachen, die in mehr als einem nationalen Zentrum als Standardsprache gebraucht werden, plurizentrischen Charakter besitzen. Das bedeutet, dass sprachliche Varianten in einem nationalen Zentrum nicht einfach Abweichungen von der Standardsprache sind, sondern dass es mehrere gleichberechtigte standardsprachliche Varietäten nebeneinander geben kann.

Entscheidend für die neuere Auffassung war die Erkenntnis, dass die Nationalstaaten eine nicht zu unterschätzende Bedeutung bei der Herausbildung und Entwicklung der Standardsprache haben. Verwaltung, Recht und Institutionen haben grossen Einfluss bei der Festsetzung des sprachlichen Standards. So werden die einzelnen Teile des politischen Systems und die verschiedenen Abläufe in Gesetzgebung und Verwaltung meist unterschiedlich bezeichnet, oder aber gleiche Bezeichnungen haben ganz unterschiedliche Bedeutungen (z. B. *Bundesrat* in Deutschland und der Schweiz). Aber auch die Sprache der Medien und Bereiche der alltäglichen Verwendung von Standardsprache sind deutlich plurizentrisch.

Standardsprachliche Variation dieser Art wurde bisher kaum systematisch untersucht und in Wörterbüchern zugänglich gemacht. Für das Deutsche liegen bisher nur Wörterbücher vor, die die Varianten in den peripheren Zentren behandeln, für die Schweiz beispielsweise Meyer 1989, für Österreich Ebner 1998 [1969]. Eine Darstellung von nur in Deutschland üblichen Varianten fehlt bisher vollständig. Die grossen Wörterbücher des Deutschen behandeln diese Varianten meist als gemeindeutsche Normalform. Im *Deutschen Wörterbuch* von Wahrig (1997, 943) steht unter dem Lemma *parken* kein Hinweis auf eine nationale oder regionale Beschränkung, *parkieren* dagegen wird klar als schweizerisch markiert. Einzig bei Varianten, die nicht in ganz Deutschland gebräuchlich sind, sind regionale Einschränkungen vermerkt. Das Wort *Sonnabend* (S. 1143) beispielsweise wird als „bes. nord- u. mitteld.“ bezeichnet.

Dieser unbefriedigenden lexikographischen Praxis soll mit einem Wörterbuch der nationalen Varianten des Deutschen begegnet werden. Darin sollen die heutigen Varianten der deutschen Standardsprache in Deutschland, Österreich und der Schweiz, aber auch in Luxemburg, Ostbelgien, Liechtenstein und Italien (Südtirol) beschrieben und erklärt und so einem breiten Publikum zugänglich gemacht werden. Das Projekt ist trinational angelegt, mit selbstständigen Arbeitsstellen in Duisburg¹, Innsbruck² und Basel³, in Zusammenarbeit mit weiteren Experten in deutschen Halbzentren (Liechtenstein, Luxemburg, Südtirol und Ostbelgien).

3. Überlegungen zum Korpus und zur Korpusauswertung eines Wörterbuchprojektes

Eine entscheidende Voraussetzung für die Erstellung eines Wörterbuches ist das Korpus, 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 Sprechens darstellen. Nun ist aber die Sprachwirklichkeit so gross und komplex, dass es keine formalen Verfahren gibt, um eine statistisch repräsentative Auswahl daraus zu treffen. Zudem ist die Berücksichtigung der mündlichen Quellen (Ansprachen, Reden, Fernsehen, Radio, Schule usw.) gegenüber schriftlichen ungleich aufwändiger, da zuerst eine Transkription erstellt werden muss. 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.

In unserem Projekt haben wir uns dafür entschieden, eine Liste von sachlichen Domänen des Sprachgebrauchs festzulegen und zu jeder dieser Domänen eine Anzahl Quellen auszuwerten. Damit sollten die wichtigsten Bereiche standardlichen Sprechens abgedeckt sein. Folgende Sprach-Domänen sind festgelegt worden: *Schweizer Romane seit 1950, Biographien, Berichte, Bildung und Erziehung, Brauchtum, Volkskunde, Geographie, Geschichte, Landeskunde, Geschäftsleben, Wirtschaft, Gesundheit, Körperpflege, Körper, Medizin, Handwerk, Handarbeit, Bau, Architektur, häusliches Leben, Wohnen, Kinder-, Jugend-, Schüler- und Studentenkultur, Kleidung, Mode, Haartracht, Kunst, Kultur, Musik, Tanz, Land- und Forstwirtschaft, Jagd, Medien, Informatik, Telekommunikation, menschliches Verhalten und Benehmen, Soziales, Nahrung, Kochkunst, Natur, Naturwissenschaft, Umwelt, öffentliche Institutionen, Post, Politik, Verwaltung, Recht, Religion, Glaube, Abergläubische Astrologie, Esoterik, Sport, Spiel, Technik, Industrie, Energie, Tourismus, Gastronomie, Verkehrswesen, öffentlicher und privater Verkehr, Schifffahrt, Flugverkehr, Wehrwesen, illustrierte Wochenzeitungen, Tageszeitungen, Amtsformulare, amtliche Mitteilungen, öffentliche Rede*. Als Quellen kommen vorzugsweise ganz aktuelle Texte in Frage, ausnahmsweise, wenn kein geeigneter neuer Text gefunden wird, können etwas ältere Texte berücksichtigt werden, allerdings keine aus der Zeit vor den 70-er Jahren. Auch bei den Romanen liegt das Schwergewicht auf Werken aus den 90-er Jahren.

In diesen Quellen werden 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 lexikaliisiert ist oder ob es sich lediglich um eine Augenblicksbildung handelt, bietet die Arbeit an einem Wörterbuch der nationalen Varianten einige zusätzliche Schwierigkeiten. Aufgenommen werden ja nicht einfach die Wörter, die in einem bestimmten Zentrum vorkommen. Vielmehr muss gesichert sein, dass ein Wort zusätzlich in einem anderen Zentrum oder in grösseren Teilen des eigenen Zentrums *nicht* vorkommt. Nötig ist also nicht nur ein positiver Test, der die Existenz eines Wortes in einem bestimmten Gebiet feststellt, sondern ebenso ein negativer Test, der das Vorkommen des Wortes an anderen Orten ausschliesst.

Diese doppelte Bedingung stellt ganz besondere Anforderungen an das methodische Vorgehen bei der Belegsammlung und an Umfang und Ausgewogenheit des Korpus. Die Wörterbuchmitarbeiter und -mitarbeiterinnen können sich nur zur Hälfte auf ihre Sprachkompetenz verlassen, nämlich bei der Feststellung, ob ein Wort in ihrem Zentrum vorkommt. Zur Beurteilung der nationalen oder geographischen Reichweite sind sie auf Hilfe von aussen angewiesen.

Bei unserem Projekt werden daher in einem Bearbeitungszentrum ausschliesslich die Quellen der jeweils beiden anderen Zentren gelesen. Dabei werden alle sprachlichen Erscheinungen angestrichen und kommentiert, die im jeweils eigenen Zentrum nicht vorkommen. Nur auf diese Weise besteht Gewähr, dass der zweite, negative Test zuverlässig durchgeführt werden kann.

Dieses an sich bewährte Verfahren kann aber dennoch nicht absolute Sicherheit bezüglich der richtigen Beurteilung von Varianten bieten. Schwierigkeiten ergeben sich einerseits daraus, dass der Wortschatz jedes Menschen beschränkt ist. Beson-

ders in Sachbereichen, die jemanden nicht besonders interessieren, bestehen auch meist deutliche Lücken im Wortschatz. Andererseits gibt es, und dies trifft in besonderer Weise auf Deutschland zu, auch Wörter, die nur in Teilbereichen eines Landes gebräuchlich sind. Es ist daher für einen Einzelnen nahezu unmöglich, bei einem bestimmten Wort mit Sicherheit auszuschliessen, dass es in einem Zentrum vorkommt.

Dazu kommt, und dies gilt insbesondere für die Schweizer BearbeiterInnen, dass die Standardsprache der anderen Zentren durch Literatur und Medien meist doch einigermassen häufig rezipiert wird, so dass es in vielen Fällen nicht mit der notwendigen Sicherheit möglich ist, das Vorkommen eines Wortes im eigenen Zentrum auszuschliessen. Man ist manchmal nicht einmal mehr ganz sicher, ob man ein Wort nicht selbst vielleicht auch brauchen würde.

Um diese Unsicherheiten auszugleichen, ist es notwendig, mit einem verlässlichen Korpus zu arbeiten, aus dem einigermassen gesicherte Angaben über das Vorkommen eines Lexems gezogen werden können. Nun ist aber der Aufbau eines 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, die wohl bei weit über 100'000 Belegen liegen. Eine solche Datenbank ist aber im vorgesehenen und finanzierten Rahmen nicht zu schaffen.

Wir haben deshalb nach Möglichkeiten gesucht, einigermassen zuverlässige Frequenzangaben zu erhalten, die die Angaben aus unserer Belegdatenbank ergänzen. Dazu ist uns die gegenwärtige Entwicklung im Internet, genauer im World-Wide-Web oder kurz WWW, gelegen gekommen. In der zweiten Hälfte der 90-er Jahre hat sich das WWW zu einem Informationsmedium entwickelt, in dem Millionen von Texten unterschiedlichster Provenienz frei zugänglich sind. Es brauchte also einzig noch ein Instrument, das diese Texte so erschliesst, dass sie als datenbankähnliche Quelle benutzt werden können.

Ein solches Instrument bieten die Suchmaschinen⁴, die einen grossen Teil der Internetseiten durch einen Index erschliessen. Als für unsere Zwecke besonders brauchbar hat sich die Suchmaschine von *AltaVista*⁵ erwiesen.

4. Die Suchmaschine AltaVista

Die Suchmaschine AltaVista wurde von Mitarbeitern einer Computerfirma entwickelt, die damit die Leistungsfähigkeit ihres neuen Prozessors für Datenbankanwendungen demonstrieren wollten. Zu diesem Zweck wurde begonnen, die auf dem WWW zugänglichen Seiten in eine Datenbank zu laden und mit einem Index zu versehen. Dieser Vorgang geschieht automatisch, Tag und Nacht wird das Internet nach Seiten abgesucht, die neu aufgenommen werden, gleichzeitig werden nicht mehr aktuelle Seiten gelöscht.⁶

Gefunden werden die Seiten von der Suchmaschine einerseits dadurch, dass Gestalter von Internetseiten diese bei AltaVista anmelden. Andererseits werden systematisch alle Links auf vorhandenen Seiten verwertet, so dass auch ein grosser Anteil nicht gemeldeter Seiten aufgenommen wird.

Im Juni 1999 waren ungefähr 330 Millionen Internetseiten bei AltaVista indiziert. Damit steht ein ausserordentlich umfangreiches Korpus zur Verfügung, das ohne weiteren Aufwand vom Forscher genutzt werden kann. Nun ist aber natürlich nur

der kleinere Teil der Internetseiten auf Deutsch verfasst. Vorherrschende Sprache ist das Englische. AltaVista bietet jedoch die Möglichkeit, alle nicht-deutschsprachigen Seiten von einer Suche auszuschliessen. Nicht nur das, man kann zusätzlich die sogenannte *Internetdomain* bestimmen, in der gesucht werden soll. Die Adressen im WWW sind in verschiedene *Domains* gegliedert. Diese kommen im letzten Teil der Internetadresse zum Ausdruck. 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 gehören⁷. Weitere Domainnamen sind z. B. *com* (für kommerzielle Anbieter), *edu* (Internetseiten des amerikanischen Hochschulnetzes), *org* (für internationale gemeinnützige Institutionen), aber auch weitere länderspezifische Domains, so auch *de* für Seiten in Deutschland und *at* für österreichische Seiten.

Dank dieser Einteilung in Domains ist es nun möglich, im AltaVista-Index gezielt nach der Frequenz von einzelnen deutschen Wörtern in den drei nationalen Zentren Österreich, Schweiz und Deutschland zu suchen. Insgesamt stehen als Korpus zurzeit bei AltaVista ungefähr 20 Millionen deutsche Seiten zur Verfügung. Dabei kann eine Seite nur ganz wenige Wörter enthalten, andere umfassen mehrere A4-Textseiten. Durch die andauernde automatisierte Indizierung ändert sich die Zahl allerdings fast ständig.

5. Die deutschen WWW-Seiten als lexikographisches Korpus

Die Hauptfrage ist nun aber: Eignet sich das Internet als Basis für lexikographische und sprachstatistische Untersuchungen? Gewöhnlich wird, wie oben in Kapitel 3 dargestellt wurde, 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. Zudem ist es mit seinen über 20 Millionen deutschen Seiten schlicht zu umfangreich. Dazu kommt, dass ständig Seiten verschwinden, andere neu hinzukommen. Der Bestand der von AltaVista indizierten Seiten ist zwischen Juni 1998 und Juni 1999 um mehrere hundert Prozent gestiegen.⁸

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

1. erwiesen wird, dass damit einigermassen zuverlässige und vor allem reproduzierbare Ergebnisse erzielt werden können;
2. wenn die Ergebnisse in einem systematischen Bezug zur Sprachwirklichkeit stehen.

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

Eine Schwierigkeit ergibt sich insofern, dass es für das Deutsche kaum sprachstatistische Arbeiten gibt, die für einen Vergleich herangezogen werden können. Der bisher einzige grössere Versuch von H. Meier (1967 [1964]) ist nach über dreissig Jahren nicht mehr ganz aktuell. Ein Vergleich ist 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 müsste 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 Tabelle 1 dargestellt.

Lexem	AltaVista Abfrageergebnisse vom 22.10.1998				AltaVista Abfrageergebnisse vom 25.2.1999				AltaVista Abfrageergebnisse vom 23.5.1999			
	A	CH	D	Gesamt	A	CH	D	Gesamt	A	CH	D	Gesamt
<i>selten</i>	4'691 8.88%	5643 10.69%	42465 80.43%	52799 100%	4'592 8.74%	5360 10.20%	42575 81.05%	52'527 100%	6'093 9.07%	7054 10.50%	54'057 80.44%	67'204 100%
<i>wollen</i>	68'700 10.13%	67'490 9.96%	54'1690 79.91%	67'7880 100%	67'410 10.16%	63'440 9.56%	53'2470 80.27%	66'3320 100%	103'009 10.70%	89'255 9.27%	77'0272 80.02%	96'2540 100%
<i>Tisch</i>	29'30 9.34%	34'20 10.90%	25'030 79.76%	31'380 100%	28'50 9.05%	33'35 10.60%	25'305 80.35%	31'493 100%	4'146 10.45%	4'645 11.71%	30'870 77.83%	39'661 100%
<i>Mensch</i>	8'064 10.27%	8'357 10.64%	62'130 79.10%	78'551 100%	8'152 10.39%	8'213 10.47%	62'063 79.13%	78'428 100%	11'170 11.33%	10'717 10.87%	76'673 77.79%	98'560 100%
<i>Baum</i>	1'843 9.33%	1'580 8.00%	16'322 82.66%	19'745 100%	1'998 9.53%	1'693 8.08%	17'271 82.39%	20'962 100%	2'604 10.49%	2'103 8.47%	20'122 81.04%	24'829 100%
<i>Kopf</i>	6'691 8.30%	8'101 10.05%	65'792 81.64%	80'584 100%	6'636 8.05%	8'063 9.78%	67'776 82.18%	82'475 100%	8'976 9.25%	9'985 10.29%	78'122 80.47%	97'086 100%
<i>soll</i>	81'040 10.53%	64'010 8.32%	62'4390 81.15%	76'9440 100%	76'470 10.15%	59'990 7.96%	61'6800 81.88%	75'3260 100%	97'952 11.17%	93'123 10.62%	68'6062 78.22%	87'137 100%
<i>schön*</i>	20'106 9.55%	21'662 10.29%	16'8835 80.17%	21'0603 100%	19'494 9.46%	20'141 9.78%	16'6394 80.76%	20'6029 100%	29'601 10.08%	31'545 10.75%	23'2380 79.17%	293'526 100%
<i>Regen</i>	1'392 7.80%	1'929 10.81 %	14'517 81.38%	17'838 100%	1'056 8.02%	1'388 10.55%	10'716 81.43%	13'160 100%	2'121 9.12%	2'425 10.43%	18'705 80.45 %	23'251 100%
<i>Computer</i>	83'050 8.24%	111'320 11.04%	81'3770 80.72%	1'008'140 100%	80'160 8.47%	103'240 10.91%	76'3242 80.63%	94'642 100%	103'786 6.55%	157'184 9.91%	1'324'495 83.54%	1'585'465 100%
Total Abs.	278'507	293'512	2'374'941	2'946'960	268'818	274'866	2'304'612	2'848'296	369'458	408'043	3'291'758	4'069'259
Total %	9.45%	9.96%	80.59%	100%	9.44%	9.65%	80.91%	100%	9.08%	10.03%	80.89%	100%

Tabelle 1: In dieser Tabelle 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 befinden sich je ca. 10% der Internetseiten, in Deutschland je ungefähr 80%. Man beachte, dass, obwohl sich die absoluten Zahlen im Lauf der Zeit stark geändert haben, die Prozentzahlen ziemlich stabil geblieben sind. So war das Wort *selten* im Oktober 1998 auf 4'691 österreichischen Seiten indiziert, im Mai 1999 bereits auf 6'093. Trotzdem haben sich die prozentualen Verhältnisse nur um 0.19% verschoben. 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 liegen für Österreich im Schnitt bei ungefähr 9.5%, für die Schweiz bei ca. 10% und für Deutschland bei 80.5%. Eine grössere Streuung gibt es einzlig beim Wort *Computer*, das in Österreich mit 6.55% im Mai 1999 deutlich weniger häufig auftrat als erwartet. Ob hier eine zufällige Streuung vorliegt oder ob in Österreich neben *Computer* das Wort *Rechner* häufiger als in den ande-

ren Zentren vorkommt, können wir vorläufig noch nicht bestimmen. Sonst liegen alle anderen Abfragen in einem sehr engen Streubereich. Wenn man zusätzlich bedenkt, dass das Korpus zwischen Oktober 1998 und Mai 1999 um ca. 38% gewachsen ist und dass zusätzlich viele Seiten verschwunden und durch andere ersetzt worden sind, haben sich die Veränderungen in ganz engen Grenzen gehalten.

Nachdem der erste Test gezeigt hatte, dass bei mehreren unterschiedlichen Lemmata auch über die zeitliche Distanz 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 werden. Es sind dies: *Maturand* (nach DUW ‘schweiz., sonst veraltet’), *Maturant* (nach DUW ‘österr.’), *Abiturient* (in DUW nicht markiert, jedoch bei Meyer unter dem Lemma *Maturand* als gar nicht üblich markiert), *allfällig* (nach DUW ‘bes. österr., schweiz.’). Die Resultate sind in Tabelle 2 dargestellt.

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

Tabelle 2: Wie in Tabelle 1 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 Tabelle 1 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 mit über 98% der Fundstellen vorwiegend in Deutschland verbreitet, *Maturand* mit fast derselben prozentualen Häufung der Belege in der Schweiz.

Die Resultate in Tabelle 2 sind eindeutig. Die Erwartungen aus dem lexikographischen Vorwissen werden bestätigt und teilweise präzisiert. *Maturand*, *Maturant* und *Abiturient* sind tatsächlich fast ausschliesslich in jeweils einem nationalen Zentrum auf Internetseiten zu finden. Die wenigen Belege, die in den jeweils anderen Zentren gefunden werden, sind meist WWW-Seiten, die Informationen über das Ursprungszentrum enthalten 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’ finden sich, entsprechend der Sprachwirklichkeit, in allen Zentren. Sie lassen sich mit AltaVista auch einzeln anschauen und überprüfen, so dass sie in der Regel erklärt werden können.

Im Unterschied zu den Angaben in DUW, wo das Wort *allfällig* als ‘bes. österr., schweiz.’ markiert ist, zeigt 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 wie auch die Sprachwirklichkeit. Die von uns durchgeföhrten 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.

6. Ausblick bzw. Perspektiven einer Internet-Lexikographie-Forschung

Der Versuch, das Internet als lexikographische Quelle zu nutzen, hat ermutigende Resultate erbracht. Das WWW ist erst wenige Jahre alt, hat sich in dieser Zeit aber bereits zu einem nicht mehr wegzudenkenden Informationsmedium entwickelt, in dem eine Vielzahl unterschiedlicher Textsorten angeboten wird. Sie reichen von persönlichen Homepages über Verwaltungs- und Gesetzestexte, wissenschaftliche Abhandlungen, Werbung und Dienstleistungsangebote, bis hin zu Zeitungs- und Zeitschriftenarchiven. Das Korpus von über 20 Millionen deutschen Texten, das sich täglich verändert, steht in einem systematischen Bezug zur verschriftlichten Sprachwirklichkeit.

Wie der Bezug genau ist, lässt sich zurzeit nicht sagen. Dies liegt daran, dass es keine aktuelle Sprachstatistik des Gegenwartsdeutschen gibt. 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 aufweisen. Dazu sind seit seiner Untersuchung mehr als 30 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 Zunahme der Internetseiten um mehrere hundert Prozent im letzten Jahr hat sich an den erzielten Resultaten kaum etwas geändert. Die Grösse und Vielfalt des Korpus garantierte seine Stabilität auch während der enormen Wachstumsphase.

Diese Ergebnisse ermutigen dazu, auch grössere Korpora von Wörtern mit dieser Methode zu analysieren. Zurzeit werden bei unserem Projekt solche Textkorpora wie einzelne Ausgaben von Tageszeitungen, Protokolle von Parlamentssitzungen und Romane zu Wortlisten verarbeitet und systematisch auf nationale Varianten

mittels automatisierter Internet-Abfrage überprüft. Dadurch wird die traditionelle Quellenauswertung auf ideale Weise ergänzt.

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.⁹ 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.

7. Anmerkungen

-
- 1 Leitung: U. Ammon, wissenschaftliche MitarbeiterInnen: B. Kellermeier, M. Schlossma-cher.
 - 2 Leitung: H. Moser, J. Ebner, wissenschaftliche MitarbeiterInnen: D. Mangott, G. Vallaster.
 - 3 Leitung: H. Bickel, H. Löffler, R. Schläpfer, wissenschaftliche MitarbeiterInnen: M. Gasser, L. Hofer, R. Schmidlin.
 - 4 Suchmaschinen sind am Internet angeschlossene Computer, die versuchen, möglichst viele der dezentral über das Internet verstreuten Seiten durch einen Index zu erschliessen. Dabei gibt es zwei grundsätzlich verschiedene Methoden: 1) Automatische Volltexterschliessung durch Suchroboter; 2) systematische Erschliessung in hierarchischen Gruppen durch ein Redaktionsteam. AltaVista gehört zur ersten Kategorie.
Die ersten Suchmaschinen sind 1994 entstanden, meist an Hochschulen in den Vereinigten Staaten. Heute sind es meist kommerzielle Anbieter, die sich über Werbung finanzieren (s. Becker 1997).
 - 5 <http://www.altavista.com>
 - 6 <http://www.altavista.digital.com/av/content/about.htm> und Becker 1997.
 - 7 Ausnahmen sind jedoch zum Bsp. die grossen multinationalen Konzerne mit Sitz in der Schweiz oder auch die internationalen gemeinnützigen Institutionen, wie z. B. das Internationale Komitee vom Roten Kreuz.
 - 8 Diese Angabe beruht auf eigenen Recherchen. Im Juni 1998 fand AltaVista das Wort *hier* z. B. auf 778'388 Internet-Seiten. Ein Jahr später verzeichnete die Suchmaschine bereits 4'541'767 Seiten für dieselbe Suchabfrage. Die Ergebnisse bei anderen Lexemen liegen ungefähr ähnlich: *Menschen* 230'886/714'112, *Frau* 174'329/593'597.
 - 9 Siehe z. B. die Zusammenstellung der Akademie der Wissenschaften zu Göttingen, <http://www.ADW-Goettingen.gwdg.de>

8. Literatur

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

- Ammon, Ulrich. 1997. „Vorüberlegungen zu einem Wörterbuch der nationalen Varianten der deutschen Sprache.“ In: Moelleken, Wolfgang W.; Weber, Peter J. (Hrsg.), *Neue Forschungsarbeiten zur Kontaktlinguistik*. Bonn: Dümmler.
- Ammon, Ulrich. 1997. *Nationale Varietäten des Deutschen*. Heidelberg: Groos. [= Studienbibliographien Sprachwissenschaft; Bd. 19]
- Becker, Ferenc. 1997. *Internet-Suchmaschinen. Funktionsweise und Beurteilung*. Elektronische Publikation, <http://machno.hbi-stuttgart.de/~beckerf/diplom>.
- Bergmann, Rolf (Hrsg.). 1988. *Probleme der Textauswahl für einen elektronischen Thesaurus*. Beiträge zum ersten Göttinger Arbeitsgespräch zur historischen deutschen Wortforschung. Leipzig.
- Clyne, Michael (Hrsg.). 1992. *Pluricentric Languages: Differing Norms in Different Nations*. Berlin, New York: de Gruyter.
- Duden. 1996. *Deutsches Universalwörterbuch [A-Z]*. 3., neu bearb. und erw. Aufl. Mannheim; Zürich [etc.] : Dudenverlag.
- Ebner, Jakob. 1998. *Wie sagt man in Österreich? Wörterbuch der österreichischen Besonderheiten*. 3., vollst. überarb. Aufl. Mannheim; Wien; Zürich: Bibliographisches Institut. [= Duden-Taschenbücher; Bd. 8]
- Hofer, Lorenz. 1999. „Ein Wörterbuch mit nationalen Varianten des Deutschen.“ In: *Sprachspiel 1*, 7-15.
- Meier, Helmut. 1967. *Deutsche Sprachstatistik*. 2., erw. und verb. Aufl. Hildesheim: Olms.
- Meyer, Kurt. 1989. *Wie sagt man in der Schweiz? Wörterbuch der schweizerischen Besonderheiten*. Mannheim; Zürich: Dudenverlag. [= Duden-Taschenbücher; Bd. 22]
- Ris, Roland. 1988. „Der schweizerische Anteil in den deutschen Grosswörterbüchern.“ In: Bergmann, Rolf (Hrsg.), *Probleme der Textauswahl für einen elektronischen Thesaurus*. Beiträge zum ersten Göttinger Arbeitsgespräch zur historischen deutschen Wortforschung. Leipzig, 113-126.
- Wahrig, Gerhard. 1997. *Deutsches Wörterbuch*. 6., neu bearb. Aufl. Gütersloh: Bertelsmann Lexikon Verlag.