Publié dans Revue Tranel (Travaux neuchâtelois de linguistique) 55, 7-23, 2011,
source qui doit être utilisée pour toute référence à ce travail

Deutsche Varietäten in Internetkorpora – eine kleine Entwicklungsgeschichte

Hans BICKEL
Universität Basel, Deutsches Seminar

The starting point for this article is the research on the dictionary Variantenwörterbuch des Deutschen, which was done between 1998 and 2004 at the Universities of Basel, Duisburg and Innsbruck. This dictionary contains all known national and regional variants of the German standard language. During the research work it became clear that the identification of such variants needed an empirical validation. Since still no useful linguistic corpora of the German language existed at that time, experiments with Internet search engines were made, first with AltaVista, later also with Google. Thanks to the frequency numbers given in the search results, it was possible to get reliable hints for national variation. Today the situation has changed. Useful linguistic corpora are now available. In this paper, it is shown on the one hand that the results of the web search engines became less reliable over the time, on the other hand that still no well-structured corpus exists, which satisfies all needs of the research on national variants.

1. Nationale und regionale Varianten der deutschen Standardsprache

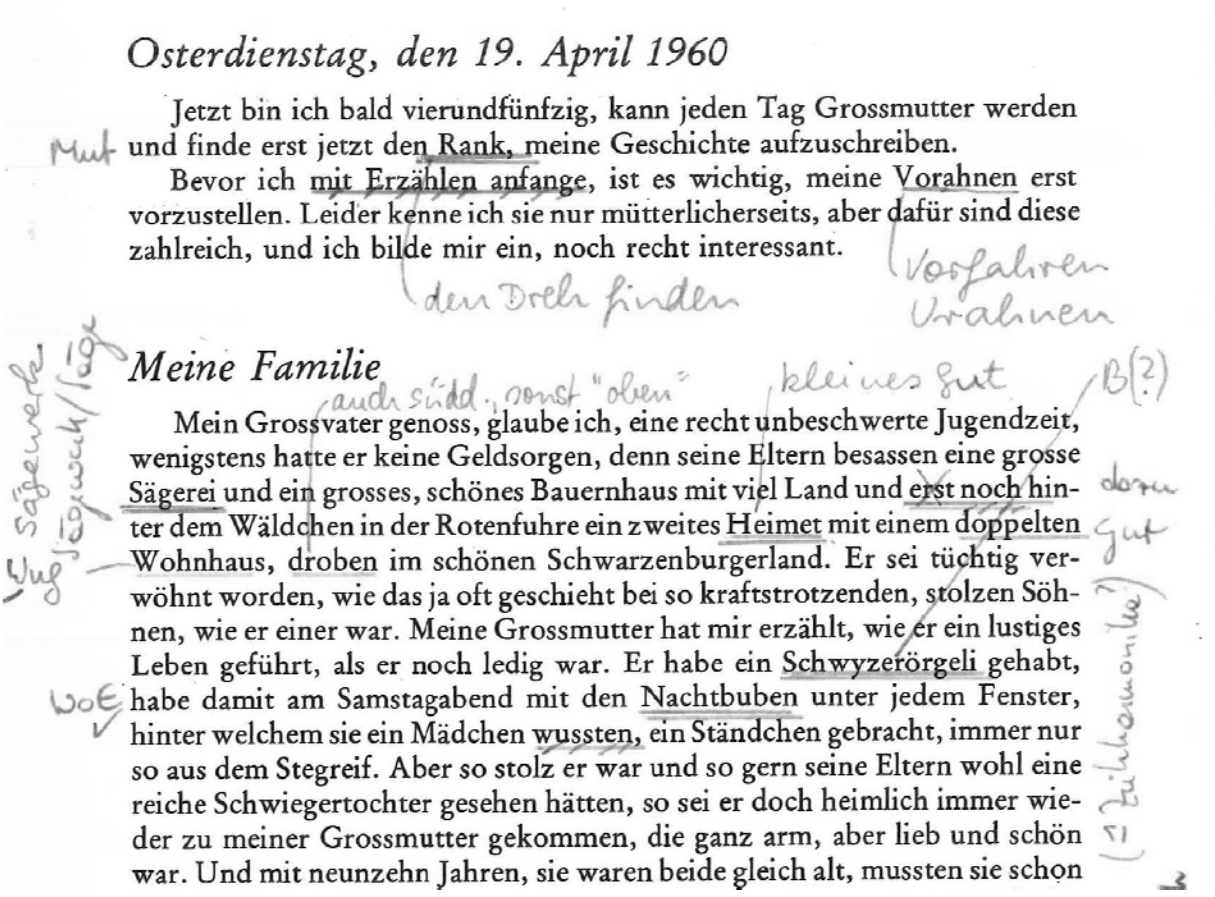
Ausgangspunkt dieser Überlegungen ist die Erforschung der nationalen und regionalen Varianten des Deutschen, an der zwischen 1998 und 2004 eine internationale Forschergruppe in Basel, Duisburg und Innsbruck gearbeitet hat und die mit der Publikation des Variantenwörterbuchs des Deutschen abgeschlossen wurde[1]. Im Variantenwörterbuch sollten alle nationalen und regionalen Varianten der deutschen Standardsprache gesammelt und entsprechend erläutert werden. Dazu mussten diejenigen standardsprachlichen Erscheinungen identifiziert werden, die nicht gemeindeutsch sind, sondern einer nationalen oder regionalen Beschränkung unterliegen. Zu diesem Zweck wurden damals Texte aus den verschiedenen nationalen Zentren (jeweils aus den Zentren, denen die jeweiligen Bearbeiter nicht selbst angehörten) gelesen, und es wurden alle Wörter sowie sprachlichen Erscheinungen aus der Lexik, Morphologie, Syntax und Pragmatik angestrichen, die aus der Sicht des eigenen Zentrums unbekannt oder auffällig waren. Anschliessend war es die Aufgabe der Mitarbeitenden der Arbeitsstellen in Innsbruck, Basel und Duisburg, die Anstreichungen, die in den Texten des eigenen Zentrums von den Mitarbeitenden aus den anderen Zentren gemacht worden waren, durchzusehen und zu bewerten. Das heisst also,

[1] Ammon & Bickel & Ebner et al. (2004).

es musste beurteilt werden, ob ein angestrichenes Wort im eigenen Zentrum als normal standardsprachlich vorkam oder ob allenfalls der Autor oder die Autorin des Textes etwas Eigenwilliges geschrieben hatte, das auch aus Sicht des eigenen Zentrums wenig geläufig oder fremdartig erschien.

Die auf den ersten Blick triviale Aufgabe war häufig schwieriger, als wir zu Anfang gedacht hatten. Es zeigte sich relativ bald, dass bei allen Mitarbeitenden ziemlich häufig Unsicherheiten bei der Identifizierung von Varianten auftraten. Dies galt vor allem dann, wenn beurteilt werden musste, ob ein sprachlicher Ausdruck gängig und typisch für das eigene Zentrum ist oder ob es sich um eine eigenwillige Formulierung eines Autors oder einer Autorin handelt. Besonders 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 im eigenen Zentrum vorkommt.

Dazu kommt, und dies gilt insbesondere für die Schweizer Bearbeiter und Bearbeiterinnen, dass die Standardsprache der anderen Zentren durch Literatur und Medien meist häufig rezipiert wird, so dass auch in der Schweiz an sich ungebräuchliche Wörter häufig mindestens passiv bekannt sind. Man ist manchmal nicht einmal mehr ganz sicher, ob man ein Wort nicht selbst auch brauchen würde.



Osterdienstag, den 19. April 1960

Jetzt bin ich bald vierundfünfzig, kann jeden Tag Grossmutter werden und finde erst jetzt den Rank, meine Geschichte aufzuschreiben.

Bevor ich mit Erzählen anfange, ist es wichtig, meine Vorahnen erst vorzustellen. Leider kenne ich sie nur mütterlicherseits, aber dafür sind diese zahlreich, und ich bilde mir ein, noch recht interessant.

Meine Familie

Mein Grossvater genoss, glaube ich, eine recht unbeschwerte Jugendzeit, wenigstens hatte er keine Geldsorgen, denn seine Eltern besassen eine grosse Sägerei und ein grosses, schönes Bauernhaus mit viel Land und erst noch hinter dem Wäldchen in der Rotenfuhre ein zweites Heimet mit einem doppelten Wohnhaus, droben im schönen Schwarzenburgerland. Er sei tüchtig verwöhnt worden, wie das ja oft geschieht bei so kraftstrotzenden, stolzen Söhnen, wie er einer war. Meine Grossmutter hat mir erzählt, wie er ein lustiges Leben geführt, als er noch ledig war. Er habe ein Schwyzerörgeli gehabt, habe damit am Samstagabend mit den Nachtbuben unter jedem Fenster, hinter welchem sie ein Mädchen wussten, ein Ständchen gebracht, immer nur so aus dem Stegreif. Aber so stolz er war und so gern seine Eltern wohl eine reiche Schwiegertochter gesehen hätten, so sei er doch heimlich immer wieder zu meiner Grossmutter gekommen, die ganz arm, aber lieb und schön war. Und mit neunzehn Jahren, sie waren beide gleich alt, mussten sie schon

Fig. 1: Ausschnitt aus der ersten Seite der Lebensbeschreibung von Rosalia Wenger mit Anstreichungen aus Österreich und Deutschland.

Fig. 1 zeigt ein Beispiel der ersten Seite der Autobiografie von Rosalia Wenger, die 1978 unter dem Titel „Rosalia G." erschienen ist. Der Text ist insofern aussergewöhnlich, als die Autorin aus einfachen Verhältnissen stammte, als Verdingkind im Emmental eine eher rudimentäre Schulbildung genoss und erst spät mit der Niederschrift ihrer Lebensgeschichte begonnen hat. Der Text ist durchsetzt mit Dialektwörtern, die die vergangene ländliche Lebenswelt möglichst authentisch evozieren sollen. Daher befinden sich in diesem Text überdurchschnittlich viele Anstreichungen.

Der Text wurde nach unserem klassischen Verfahren in Duisburg und Innsbruck gelesen und mit den handschriftlichen Anmerkungen an die Basler Arbeitsstelle zur Auswertung zurückgeschickt, damit die standardsprachlichen Helvetismen identifiziert und inventarisiert werden konnten. Hier musste dann beurteilt werden, ob beispielsweise eine Wendung wie *den Rank finden* tatsächlich ein reiner Helvetismus ist, ob sie möglicherweise nur mundartlich gebraucht wird oder ob sie auch in anderen standardsprachlichen Texten vorkommt.

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, bietet die Arbeit an einem Wörterbuch der nationalen Varianten

einige zusätzliche Schwierigkeiten. Aufgenommen wurden 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.

Um solche Fragen zu prüfen, kam ziemlich bald das Bedürfnis auf, das nur teilweise zuverlässige Sprachgefühl empirisch abzusichern. Dazu boten sich die Methoden der Korpuslinguistik an.

2. Das Web als Korpus 1998/1999

1998 gab es als einziges grösseres linguistisches Korpus das IdS-Korpus[2] in Mannheim. Eine Auswertung nach Nationen oder Regionen war aber damals nicht vorgesehen und nicht möglich. Es kam daher für diese Fragestellung nicht in Frage, da man jeweils alle Belege hätte durchsehen müssen, um sich ein Bild von der Verteilung machen zu können. Somit war das Web die einzige Ressource, die für nach Nationen getrennte Recherchen zur Verfügung stand. Die führende Suchmaschine damals war AltaVista. Wir haben daher Versuche unternommen, um herauszufinden ob sich AltaVista für die vorgesehene empirische Abstützung der durch die Lektüre gewonnenen Varianten eignete.

Voraussetzung für den Einsatz von AltaVista war, dass die Resultate nachvollziehbar, reproduzierbar und konsistent waren. Wir haben zu diesem Zweck eine Liste von zehn willkürlich ausgewählten Wörtern erstellt, von denen wir annahmen, dass sie gemeindeutsch sind und keiner nennenswerten regionalen oder nationalen Einschränkung unterliegen.

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

[2] https://cosmas2.ids-mannheim.de/cosmas2-web/ (25. 9. 2011)

	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%**

Fig. 2: Absolute und relative Verteilung ausgewählter Lexeme auf von AltaVista indizierten Webseiten in Österreich, der Schweiz und Deutschland im Oktober 1998.

AltaVista Abfrageergebnisse vom 23.5.1999				
Lexem	**A**	**CH**	**D**	**Gesamt**
Selten	6'093	7054	54'057	67'204
	9.07%	10.50%	80.44%	100.00%
wollen	103'009	89259	770'272	962'540
	10.70%	9.27%	80.02%	100.00%
Tisch	4'146	4645	30'870	39'661
	10.45%	11.71%	77.83%	100.00%
Mensch	11'170	10717	76'673	98'560
	11.33%	10.87%	77.79%	100.00%
Baum	2'604	2103	20'122	24'829
	10.49%	8.47%	81.04%	100.00%
Kopf	8'976	9988	78'122	97'086
	9.25%	10.29%	80.47%	100.00%
soll	97'952	93123	686'062	877'137
	11.17%	10.62%	78.22%	100.00%
schön*	29'601	31545	232'380	293'526
	10.08%	10.75%	79.17%	100.00%
Regen	29'601	31545	232'380	293'526
	10.08%	10.75%	79.17%	100.00%
Computer	103'786	157184	1'324'495	1'585'465
	6.55%	9.91%	83.54%	100.00%
Total Abs.	**396'938**	**437'163**	**3'505'433**	**4'339'534**
Total %	**9.15%**	**10.07%**	**80.78%**	**100.00%**

Fig. 3: Wiederholung der Abfrage aus Fig. 2 sieben Monate später.

Die Ergebnisse bei diesen zehn ausgewählten Wörtern haben deutlich gezeigt, dass bei national nicht markierten Wörtern durchaus vergleichbare Resultate zustande kommen. Die prozentualen Werte lagen 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 einzig beim Wort *Computer*, das in Österreich mit 6.55% im Mai 1999 deutlich weniger häufig auftrat als

erwartet. Hier war wohl die Wahl des Lexems etwas naiv, gibt es doch neben *Computer* auch noch den Terminus *Rechner*, und es ist durchaus möglich, dass hier nationale Präferenzen in die eine oder andere Richtung vorliegen. Alle anderen Abfragen liegen jedoch in einem sehr engen Streuungsbereich. 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.

Als zweiten Schritt haben wir drei Wörter abgefragt, die in der Lexikografie als typische und spezifische nationale Varianten galten, und zusätzlich eine unspezifische[3] Variante, die sowohl in Österreich wie in der Schweiz geläufig sein sollte.

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	9'021
	26.26%	70.23%	3.51%	100.00%

Fig. 4: Absolute und relative Abfrageergebnisse bekannter spezifischer und unspezifischer nationaler Varianten mit der Suchmaschine AltaVista 1998.

Die Abfrageergebnisse für bekannte Varianten haben auf den ersten Blick gezeigt, dass sich die nationale Verteilung auf von AltaVista indizierten Seiten deutlich manifestiert. Ein Wert von beispielsweise über 98% für den bekannten Helvetismus *Maturand* hat die Erwartung einer empirischen Absicherung nationaler Varianten eindrücklich bestätigt. Auch die unspezifische Variante *allfällig* hatte eine gegenüber der normalen Verteilung stark abweichende Frequenz auf österreichischen und noch wesentlich stärker auf schweizerischen Internetseiten. Als These liess sich daraus ableiten, dass *allfällig* eine österreichisch-schweizerische Variante darstellt, die jedoch in der Schweiz weit geläufiger ist als in Österreich.

Als Konsequenz aus den ermutigenden Erfahrungen mit dem Vergleich zwischen bekannten gemeindeutschen Lexemen und nationalen Varianten wurde eine Datenbank aufgebaut, in der systematisch und automatisiert ganze Wortlisten abgefragt wurden. Auch Phraseologismen wurden auf diese Weise überprüft. Wenn eine Wortform oder eine Wendung in Österreich

[3] Zu den Begriffen *spezifische* und *unspezifische Variante* s. Ammon (1995: 71).

oder der Schweiz mehr als 20% oder in Deutschland mehr als 90% erreichte, wurde diese in der Datenbank automatisch als "Verdachtsfall" für nationale oder regionale Variation markiert.

Ausdruck	AT	CH	DE	%	%	%	
den-Rank-finden	0	2	0	0%	100%	0%	Helv
den-Rank-gefunden	0	1	0	0%	100%	0%	Helv

Fig. 5: Beispiel für den Phraseologismus *den Rank finden,* wie er in der internen Forschungsdatenbank gespeichert wurde.

Phraseologismen wie beispielsweise *den Rank finden* waren auf Internetseiten naturgemäss weit schwächer vertreten als einzelne Wortformen. In diesem Fall wurde die Wendung nur gerade drei Mal gefunden. In der Regel stellten sich aber auch hier verwertbare Resultate ein. Neben der reinen Frequenz interessierte natürlich auch die Art von Texten, in denen seltenere Varianten vorkamen.

Im Lauf der Arbeit hat sich Google als führende Suchmaschine etabliert, die relative Verteilung der Abfrageergebnisse blieb jedoch trotz enormen Wachstums und entsprechender Vervielfachung der absoluten Zahlen weitgehend vergleichbar.

Auch zur Beleggewinnung[4] waren die Suchresultate häufig brauchbar, wobei allerdings die Zeitungsarchive und die mit traditionellen Exzerpten gewonnenen literarischen Belege häufig besser, d. h. sprechender waren, während die Internetbelege bereits damals repetitiver und manchmal auch belangloser waren.

Fazit: Die Erfahrung in den neunziger Jahren liess kaum ein Bedürfnis für ein eigentliches linguistisches Korpus aufkeimen. Die für die Erforschung nationaler Varianten notwendige empirische Abstützung konnte mit Hilfe von Web-Abfragen bei den gängigen Suchmaschinen zuverlässig bewerkstelligt werden.

3. Web als Internetkorpus 2011

Da für die nächste Zukunft eine Erweiterung und Aktualisierung des Variantenwörterbuchs vorgesehen ist, ist nun der richtige Zeitpunkt gekommen, um noch einmal neu zu evaluieren, ob das Web immer noch zur Bestimmung nationaler Varianten taugt oder ob die in der Zwischenzeit entstandenen und weiterentwickelten linguistischen Korpora für diese Art der For-

[4] Beispielsweise wird der Helvetismus *Trute* im Variantenwörterbuch mit dem folgenden Beleg illustriert: "*Gemästete Truten sind krank, aggressiv und viel zu schwer* (Verein gegen Tierfabriken Schweiz, 1999, Internet)"

schung brauchbarer sind und in welchem Verhältnis die Resultate aus dem Web zu denen aus den Korpora stehen.

Der Suchmaschinenmarkt wird gegenwärtig fast ausschliesslich von Google dominiert, auch wenn AltaVista nominell weiter existiert. Mit *bing* von Microsoft ist zwar eine weitere Suchmaschine mit grossen Ambitionen entstanden. Ihr Marktanteil bewegt sich jedoch in Deutschland noch im tiefen einstelligen Prozentbereich[5]. Wir beschränken uns daher in den weiteren Ausführungen auf die Suchmaschine von Google.

Was gegenüber der früheren Untersuchung gleich geblieben ist, ist die weitgehende Intransparenz der Technik hinter den Suchmaschinen. Der verwendete Algorithmus ist geheim, um so Manipulationen von Suchresultaten durch Webseitenbetreiber zu erschweren. Geändert haben sich aber die Grössenordnungen der Fundstellen, die bei häufig vorkommenden Lemmata nicht mehr exakt, sondern um mehrere Stellen gerundet anzeigt werden.

Google Abfrageergebnisse vom 21.2.2011				
Lexem	**A**	**CH**	**D**	**Gesamt**
selten	8'620'000	6'410'000	56'300'000	71'330'000
	12.08%	8.99%	78.93%	100%
wollen	47'600'000	34'700'000	230'000'000	312'300'000
	15.24%	11.11%	73.65%	100%
Tisch	8'530'000	6'860'000	47'600'000	62'990'000
	13.54%	10.89%	75.57%	100%
Mensch	13'000'000	10'600'000	75'000'000	98'600'000
	13.18%	10.75%	76.06%	100%
Baum	3'840'000	2'410'000	21'700'000	27'950'000
	13.74%	8.62%	77.64%	100%
Kopf	13'400'000	10'400'000	64'200'000	88'000'000
	15.23%	11.82%	72.95%	100%
soll	67'200'000	56'000'000	314'000'000	437'200'000
	15.37%	12.81%	71.82%	100%
schön	19'200'000	13'900'000	135'000'000	168'100'000
	11.42%	8.27%	80.31%	100%
Regen	6'520'000	8'600'000	44'200'000	59'320'000
	10.99%	14.50%	74.51%	100%
Computer	38'600'000	33'300'000	277'000'000	348'900'000
	11.06%	9.54%	79.39%	100%
Total Abs.	**226'510'000**	**183'180'000**	**1'265'000'000**	**1'674'690'000**
Total %	**13.53%**	**10.94%**	**75.54%**	**100%**

Fig. 6: Abfrageergebnis für gemeindeutsche Lemmata von Google (Stand: Februar 2011).

[5] Gemäss der Webseite http://www.luna-park.de/home/internet-fakten/suchmaschinen-marktanteile.html liegt der Marktanteil von *bing* in Deutschland bei 1.7%, während Google einen Anteil von 93.9% hat (Stand: März 2011).

Eine Abfrage der Liste der gemeindeutschen Lemmata, wie sie in den 1990er-Jahren verwendet wurde, zeigt, dass sich die neuen Ergebnisse noch innerhalb einer bestimmten Bandbreite bewegen. Aber es kommen innerhalb der Reihe einer Nation doch nennenswerte Abweichungen von über 6% vor, die nicht wirklich nachvollziehbar und erklärbar scheinen. So erhalten wir beispielsweise für das Lemma *wollen* auf deutschländischen Webseiten einen Wert von 73.75%, für *schön* einen solchen von 80.31%. Auch bei der gezielten Suche nach nationalen Varianten hat sich das relative Ergebnis im Vergleich mit den früheren Erfahrungen deutlich verändert.

Google Abfrageergebnisse vom 13.3.2011				
Lexem	**A**	**CH**	**D**	**Gesamt**
Maturand	130	9'110	1'320	10'560
	1.23%	86.27%	12.50%	100%
Maturant	59'000	9410	15'900	84'310
	69.98%	11.16%	18.86%	100%
Abiturient	10'300	7190	753'000	770'490
	1.34%	0.93%	97.73%	100%
allfällig	52'200	128'000	37'100	217'300
	24.02%	58.90%	17.07%	100.00%

Fig. 7: Ergebnisse der Google-Suche nach nationaler Varianten (Stand: März 2011).

Vereinfachend kann man sagen, dass die Kontamination der Ergebnisse zugenommen hat. Diese liefern zwar immer noch Hinweise auf nationale Variation. Sie sind also durchaus interpretierbar. Aber die Bedeutung der nationalen Domains *at/ch/de* im Web nimmt ab, einerseits weil es neue Toplevel-Domains wie *info, name* oder *travel* gibt, zum anderen weil immer häufiger identische Inhalte tel quel unter den verschiedenen Länderdomains angeboten werden. So nimmt die Brauchbarkeit des Webs als linguistisches Korpus insgesamt ab. Man findet buchstäblich für jedes Lemma in jeder Nation eine grosse Trefferzahl. So ist beispielsweise der Austriazismus *Maturant* in der Schweiz in absoluten Zahlen häufiger nachweisbar als der spezifische Helvetismus *Maturand.* Die Entwicklung hat also nicht etwa zu einem noch zuverlässigeren Resultat geführt, sondern vielmehr zu einer Verschlechterung.

4. Vergleich mit linguistischen Korpora

Zum Vergleich sollen die Ergebnisse der folgenden Korpora herangezogen werden:

- Wortschatz Leipzig (wortschatz.uni-leipzig.de)
- DWDS Berlin (www.dwds.de)
- Cosmas Mannheim (cosmas2.ids-mannheim.de/cosmas2-web)

- Korpus-C4 Berlin, Basel, Wien und Bozen (www.korpus-c4.org) bzw. Schweizer Textkorpus (dwds.ch)

Die Ausführungen zu den ersten beiden Korpora sind kurz: Beide erlauben keine Filterung der Ergebnisse nach nationalen, regionalen oder sonstigen geografischen Kriterien. *Wortschatz Leipzig* erlaubt überhaupt keine Differenzierungen. Das Korpus kann bei unserer Fragestellung höchstens für die Beleggewinnung genutzt werden. Das DWDS-Korpus in Berlin, das sich vom Anspruch her auch als ein Korpus der gesamten deutschen Sprache versteht und daher auch Texte aus Österreich und der Schweiz enthält, erlaubt zwar Filterungen, z. B. nach Dekaden, nach Textsorte oder Autor, nicht jedoch nach Entstehungsort eines Textes. Beide Korpora kommen wohl nicht zufällig aus der Computerlinguistik und sind stärker technisch als philologisch motiviert. Anders sieht es beim *Archiv der geschriebenen Sprache* von Cosmas aus.

	Cosmas II			
Lexem	**A**	**CH**	**D**	**Gesamt**
selten	33'172	22'271	121'223	176'666
	18.78%	12.61%	68.62%	100%
wollen	258'866	130'012	822'651	1'211'529
	21.37%	10.73%	67.90%	100%
Tisch	32'328	20'833	110'203	163'364
	19.79%	12.75%	67.46%	100%
Mensch	41'565	25'567	120'641	187'773
	22.14%	13.62%	64.25%	100%
Baum	23'294	7'681	73'100	104'075
	22.38%	7.38%	70.24%	100%
Kopf	82'506	35'698	203'686	321'890
	25.63%	11.09%	63.28%	100%
sollen	270'776	139'094	775'546	1'185'416
	22.84%	11.73%	65.42%	100%
schön	849'855	374'350	2'413'808	3'638'013
	23.36%	10.29%	66.35%	100%
Regen	30'759	16'721	95'724	143'204
	21.48%	11.68%	66.84%	100%
Computer	22'006	13'257	82'158	117'421
	18.74%	11.29%	69.97%	100%
Total Abs.	**1'645'127**	**785'484**	**4'818'740**	**7'249'351**
Total %	**22.69%**	**10.84%**	**66.47%**	**100%**

Fig. 8: Absolute und relative Verteilung gemeindeutscher Lexeme im Cosmas-Korpus *Archiv der geschriebenen Sprache* (Stand Frühjahr 2011).

Die relative Verteilung der gemeindeutschen Lexeme weicht von derjenigen Googles natürlich ab. Der Aufbau des Archivs der geschriebenen Sprache bei Cosmas orientiert sich nicht an der jeweiligen Grösse der Sprachgebiete. So sind österreichische Texte proportional doppelt so häufig vertreten wie bei Google. Die Streuung ist ebenfalls einiges grösser als bei den ersten Suchmaschinenergebnissen aus den 1990er Jahren.

Wenn man die Liste der nationalen Varianten in Cosmas eingibt, wird aber immerhin klar, dass das *Archiv der geschriebenen Sprache* im Hinblick auf die Länderzuordnung der Texte ganz offenbar viel weniger mit Texten aus anderen Ländern kontaminiert ist als das heutige Google-Korpus. Wir erhalten für alle drei eindeutigen Varianten Werte von über 98%.

Cosmas II				
Lexem	**A**	**CH**	**D**	**Gesamt**
Maturand	0	172	3	175
	0.00%	98.29%	1.71%	100%
Maturant	590	0	4	594
	99.33%	0.00%	0.67%	100%
Abiturient	14	19	2'095	2'128
	0.66%	0.89%	98.45%	100%
allfällig	144	520	13	677
	21.27%	76.81%	1.92%	100%

Fig. 9: Absolute und relative Abfrageergebnisse bekannter spezifischer und unspezifischer nationaler Varianten im Cosmas-Korpus *Archiv der geschriebenen Sprache* (Stand Frühjahr 2011).

Das *Archiv der geschriebenen Sprache* von Cosmas lässt zwar noch Wünsche offen, insbesondere in Bezug auf Berücksichtigung einer möglichst grossen Vielfalt unterschiedlicher Textsorten, aber auch in Bezug auf eine grössere diachronische Tiefe und eine geografischen Binnendifferenzierung Deutschlands. Es bietet aber mindestens eine gute Basis für die Identifizierung der nationalen Varianten auf der synchronen Ebene.

Ein etwas anderer Weg wurde bei der Erarbeitung des Korpus-C4 eingeschlagen. Dieses Korpus ist eine gemeinsame Initiative des *Digitalen Wörterbuchs der deutschen Sprache des 20. Jahrhunderts (DWDS),* des *Austrian Academy Corpus (AAC)*, des *Korpus Südtirol* und des *Schweizer Textkorpus (CHTK)*[6].

Das Korpus unterscheidet sich insofern von anderen Korpora, als von Anfang an ein strukturierter Aufbau vorgesehen war. Es sollte den Wortschatz des gesamten 20. Jahrhunderts möglichst breit erfassen. Das Korpus be-

[6] Einen Überblick über Aufbau, Ziele und Methoden des C4-Korpus findet sich bei Bickel, Hans & Gasser, Markus & Häcki Buhofer, Annelies et al. (2009).

steht darum aus gedruckten und maschinengeschriebenen Texten jeglicher Produktions- und Publikationsform, möglichst ausgewogen zusammengestellt nach zeitlichen und inhaltlich-sachlichen Kriterien. Konkret wurden die Texte ausgewählt nach Textsorte (also nach einen formalen Kriterium, d. i. Belletristik, Sachtexte, journalistische Prosa und Gebrauchstexte), nach Jahrhundertviertel (zeitliches Kriterium in Vierteljahrhundert-Schritten) und nach Sachgruppe (inhaltliches Kriterium, Grundlage bildete die Schlagwortnormdatei SWD mit 36 Oberkategorien).

Der Aufbau eines solchen Korpus ist natürlich um ein vielfaches aufwändiger als der eines opportunistischen Korpus, das in der Regel ja nur bereits anderswo digitalisierte oder bereits digital existierende Texte enthält. Zudem hatten die verschiedenen daran beteiligten Arbeitsstellen unterschiedliche Rahmenbedingungen. Das hat dazu geführt, dass in Bezug auf die Grösse nicht alle Zielvorgaben erfüllt werden konnten, hauptsächlich was die Anzahl der Textwörter der einzelnen Teilkorpora anbelangt. Im März 2011 setzte sich das Korpus aus 20 Mio. Textwörtern des *DWDS*, 1.7 Mio. Textwörtern des *Korpus Südtirol*, 4.1 Mio. Textwörtern des *AAC* und 20 Mio. Textwörtern des *CHTK* zusammen.

Die unterschiedlichen Grössen der Teilkorpora führen zu entsprechend abweichenden Ergebnissen. Kommt dazu, dass durch die relative Kleinheit der Korpora relativ grosse Streuungen entstehen, so dass eine statistische Auswertung wenig sinnvoll erscheint (vgl. Fig. 10).

	C4-Korpus				
Lexem	A	CH	D	s-Tir	Gesamt
selten	590	3'903	2'955	197	7'645
	7.72%	51.05%	38.65%	2.58%	100%
wollen	5'949	26'979	32'184	2'109	67'221
	8.85%	40.13%	47.88%	3.14%	100%
Tisch	631	3'030	3'570	153	7'384
	8.55%	41.03%	48.35%	2.07%	100%
Mensch	4'891	17'436	16'818	1'153	40'298
	12.14%	43.27%	41.73%	2.86%	100%
Baum	489	2'772	2'404	74	5'739
	8.52%	48.30%	41.89%	1.29%	100%
Kopf	1'058	5'915	7'139	385	14'497
	7.30%	40.80%	49.24%	2.66%	100%
sollen	5'050	30'584	34'087	3'152	72'873
	6.93%	41.97%	46.78%	4.33%	100%
schön	2'287	9'860	6'469	407	19'023
	12.02%	51.83%	34.01%	2.14%	100%

Regen	11	29	748	4	792
	1.39%	3.66%	94.44%	0.51%	100%
Computer	0	1'125	217	36	1'378
	0.00%	81.64%	15.75%	2.61%	100%
Total Abs.	**20'956**	**101'633**	**106'591**	**7'670**	**236'850**
Total %	**8.85%**	**42.91%**	**45.00%**	**3.24%**	**100%**

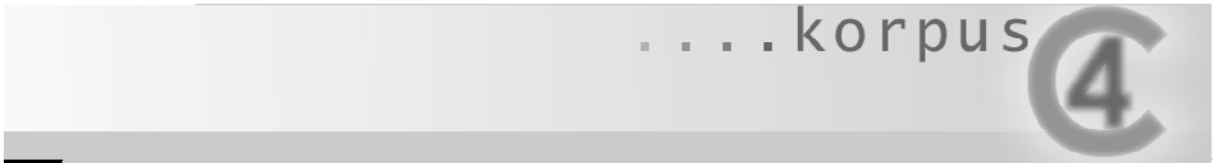
Fig. 10: Absolute und relative Verteilung gemeindeutscher Lexeme im C4-Korpus. Die relative Kleinheit und die unausgewogene Verteilung zwischen den verschiedenen Projektstellen erlaubt noch keine statistische Auswertung (Stand Frühjahr 2011).

Ein brauchbareres Bild dagegen gibt die Abfrage der bekannten nationalen Varianten. Ich beschränke mich aber hier nur auf die Teilkorpora aus Deutschland und der Schweiz, weil erst von diesen beiden Korpora je 20 Mio. Textwörter vorliegen.

C4-Korpus			
Lexem	CH	D	Gesamt
Maturand	6	0	6
	100.00%	0.00%	100%
Maturant	0	0	0
	0.00%	0.00%	0%
Abiturient	24	54	78
	30.77%	69.23%	100%
allfällig	657	8	665
	98.80%	1.20%	100%

Fig. 11: Absolute und relative Abfrageergebnisse bekannter spezifischer und unspezifischer nationaler Varianten im C4-Korpus. Berücksichtigt sind nur schweizerische und deutsche Texte (Stand Frühjahr 2011).

Wenn man die ersten beiden Stichwörter, *Maturand* und *Maturant* anschaut, werden die Erwartungen vollständig erfüllt. Der Helvetismus *Maturand* kommt nur im schweizerischen Teilkorpus vor, der Austriazismus *Maturant* fehlt im schweizerischen und im deutschen Teilkorpus. Auch beim Frequenzhelvetismus *allfällig* stammen praktisch alle Belege aus Schweizer Texten. Nicht den Erwartungen entspricht das Resultat beim Teutonismus *Abiturient*. Dies ist aber nicht auf Kontamination der Schweizer Resultate durch deutschländische Texte des C4-Korpus zurückzuführen, sondern liegt an der diachronen Ausrichtung des Korpus. In der ersten Hälfte des 20. Jhs. war der Terminus *Abiturient* auch in der Schweiz üblich, der jüngste Beleg stammt von 1961. Anschliessend ist diese Bezeichnung durch den Helvetismus *Maturand* abgelöst worden, wie ein Blick auf Fig. 11 zeigt.



CH	1913	... Die Kurse waren bestimmt für katholische Universitätsstudenten, **Abiturienten**, jüngere Akademiker, Lehrer, Studentinnen und sonstige Gebildete. ...
CH	1915	... und war weit von ihnen entfernt. Am Abend, auf dem Gartenfest der **Abiturienten** , saß er, in sich gekehrt und unzufrieden mit dem nörgeligen Still ...
CH	1919	... heiß und traumumsponnen ersehnte. In bekränzten Wagen fuhren die **Abiturienten** durch die Stadt und hinaus aufs Land, wo das Wirtshaus am Rande ...
CH	1919	... zu selbständigem Arbeiten mitbringen; bisher haben die meisten **Abiturienten** der Gymnasien erklärt, daß sie „ bis zu oberst genug haben von der ...
CH	1920	... Herr Dr. A. Mantel, zweiter Erziehungssekretär. Den sämtlichen **Abiturienten** konnte das Zeugnis der Reife erteilt werden. Sie widmen sich ...
CH	1921	... " für Aerzte, Zahnärzte, Apotheker und Tierärzte, wobei für die **Abiturienten** der Industrieschule noch eine Nachprüfung im Lateinischen ...
CH	1921	... der Präsident desselben, Herr Dekan Meier. Die Namen der **Abiturienten** sind: 1) Dehli Helge, Genua (Kaufmann, Universität Zürich). ...
CH	1927	... werden durch das Studium des letzten Bildes, der Zahlen der **Abiturienten** von Gymnasium und Realschule Basel. Hier liegt das Maximum in den ...
CH	1927	... definitive Anstellung, Heirat usw. Die Schicksale unserer **Abiturienten** von 1916 lieferten uns ein anschauliches Bild der wirtschaftlichen ...
CH	1927	... und Physik und daneben Sprachkenntnisse, welche zur Wahl von **Abiturienten** Anlaß geben. Auch hier verdrängt der Halbakademiker keineswegs ...
CH	1930	... und der Berichterstatter. Im Juni 1928 wurde an alle früheren **Abiturienten** und eine Reihe weiterer früherer Schüler, deren Adresse ausfindig ...
CH	1943	... Jahren durchgeführte Statistik zeigte, daß mehr als 50% der **Abiturienten** unserer Kollegien sich dem Theologiestudium zuwenden. Diese ...
CH	1946	... Bildungswesens; Gedanken über Nach- und Aufbauschulung für **Abiturienten** und Begabte. » Alle Schriften sind erhältlich im ...
CH	1953	... dass, wenn sie einmal geschaffen ist und sich bewährt hat, ihren **Abiturienten** selbstverständlich die gleichen Möglichkeiten einzuräumen sind wie ...
CH	1953	... Abbröckeln wirkungsvoll bekämpft werden. Wir schlagen vor, den **Abiturienten** der Abschlussklasse dieser Schule den Zugang zu öffnen zur ...
CH	1961	... uns weit verbreitet. Genauere Forschungen z.B. über den Bedarf an **Abiturienten** in fünf Jähren, Untersuchungen über die Auswirkungen verschiedener ...
CH	1964	... angetan hatten. Es war ein geladener Gast, neunzehnjähriger **Maturand** . Er hieß Manfred. ...
CH	1972	... minimale Studiendauer, Prüfungsfächer, Studienkosten usw. sei der **Maturand** auf die für jede Fakultät erschienenen< Akademischen ...
CH	1972	... stehen, der weder Befriedigung noch Erfolg zu bringen vermag. Der **Maturand** , der sich dem Jusstudium zuwenden will, sollte die Freude ...

CH	1972	... An wenige Wissensgebiete wie gerade die Jurisprudenz tritt der **Maturand** so unvorbereitet heran; er braucht deshalb genügend Zeit, um mit ...
CH	1988	... von Neumann, 1903 in Budapest geboren, machte sich bereits als **Maturand** einen Namen mit mathematischen Arbeiten. Er studierte ...

Fig. 12: Belege aus dem C4-Korpus für die Lexeme *Abiturient* und *Maturand* in Schweizer Texten. Man sieht, dass sich *Maturand* zeitlich an *Abiturient* anschliesst.

Auch wenn mit dem Korpus-C4 noch keine aussagekräftigen Statistiken über die regionale Verteilung eines Lexems gemacht werden können, bietet es doch wertvolle Einblicke in diachrone Prozesse von sprachlichen Varianten. Es ist bisher das einzige deutschsprachige Korpus, das sowohl chronologische wie auch nationale bzw. regionale Filterung erlaubt.

5. Schlussfolgerungen

Die Tests haben gezeigt, dass für die Erforschung nationaler und regionaler Varianten der deutschen Standardsprache noch keine befriedigende Korpus-Lösung existiert. In den vergangenen dreizehn Jahren ist zwar in der Korpuslinguistik viel geschehen, aber wir sind noch weit entfernt von einem idealen Korpus als universellem linguistischem Arbeitsinstrument.

Aus den bisherigen Erfahrungen ziehe ich die folgenden Schlussfolgerungen:

1. Es braucht linguistische Korpora. Anders als in der Pionierphase nimmt die Brauchbarkeit der Web-Suchmaschinen für linguistische Fragestellungen ab. Ihr eigentlicher Zweck, die Auffindung von Informationen, führt auf der linguistischen Seite zu disparateren Ergebnissen. Zudem sind kaum komplexe Suchanfragen mit regulären Ausdrücken möglich.
2. Das A und das O der linguistischen Korpora sind die Metainformationen zu den einzelnen Texten. Die Texte sollten in den Korpora mindestens mit den folgenden Kriterien versehen sein:
   - Autor
   - Geografische Verortung des Textes (z. B. durch Bestimmung des Lebensmittelpunktes des Autors, seines Wohnorts zur Zeit der Textpublikation, des Publikationsorts des Textes).
   - Textsorte
   - Inhaltskategorien, wie sie beispielsweise vom Schweizer Textkorpus verwendet wurden
3. Ein linguistisches Korpus sollte eine ansehnliche Grösse haben. 20 Mio. Textwörter, wie sie im Schweizer Textkorpus enthalten sind, sind für viele Fragestellungen zu gering.

4. Unabdingbar ist auch ein systematisch strukturierter Aufbau, so dass möglichst alle Textsorten und Fachgebiete entsprechend ihrer Bedeutung berücksichtigt werden.
5. Wünschenswert ist ferner eine möglichst umfangreiche zeitliche Abdeckung, so dass auch diachrone Fragen angegangen werden können (wünschenswert, aber wohl noch für lange Zeit Zukunftsmusik, wäre natürlich ein Korpus, das die gesamte neuhochdeutsche Periode von ungefähr 1650 bis heute abdeckt).
6. PoS-Tagging ist unerlässlich, hilfreich ist auch ein Name-Tagging, um die Kontamination möglichst klein zu halten.
7. Es stellt sich aber auch die Frage, warum in den letzten zwanzig Jahren, in denen die Bedeutung der Korpuslinguistik letztlich unbestritten war, kein Korpus entstanden ist, das den genannten Kriterien entspricht. Zwar sind einzelne grössere und kleinere Korpora entstanden. Keines davon ist jedoch entweder gross oder universell genug, dass damit die meisten linguistischen Fragestellungen beantwortet werden könnten[7]. Wünschenswert ist darum der Aufbau eines einfach zu benutzenden, universellen Korpus. Dies braucht Zeit und Geld. Deshalb stehen in erste Linie die Wissenschaftsgemeinschaft und die Förderinstitutionen in der Pflicht. Die Fokussierung auf kurzfristige Erfolge, auf Innovation und Technik ist nicht grundsätzlich falsch. Aber es gibt wissenschaftliche Vorhaben, die in drei bis fünf Jahren nicht zu bewältigen sind. Dazu gehören beispielsweise Wörterbücher, Sprachatlanten, dazu gehört aber auch die Erstellung von Korpora. Es gibt eine gewisse Hoffnung, dass mit der Einrichtung von Forschungsinfrastrukturen, die voraussichtlich bei den Akademien angesiedelt sein werden, eine grössere Bereitschaft zu einer langfristigen Finanzierung entsteht. Allerdings sind noch keine entsprechenden Mittel gesprochen.

[7] Für die Erarbeitung der Variantengrammatik, die zur Zeit unter der Leitung von Christa Dürscheid an der Universität Zürich in Angriff genommen wird, soll aus diesem Grund ein weiteres, auf die spezifischen Bedürfnisse des Projekts zugeschnittenes Korpus entstehen. Siehe den Projektbeschrieb unter http://www.variantengrammatik.net (2. 12. 2011).

Bibliografie

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

Ammon, Ulrich & Bickel, Hans & Ebner, Jakob et al. (2004): Variantenwörterbuch des Deutschen. Die Standardsprache in Österreich, der Schweiz und Deutschland sowie in Liechtenstein, Luxemburg, Ostbelgien und Südtirol. Berlin (de Gruyter).

Belica, Cyril & Keibel, Holger & Kupietz, Marc & Perkuhn, Rainer (2007): Web as Corpus: Kooperation mit der Universität Bologna. In: Sprachreport Sonderheft/März 2007, 21-25.

Bickel, Hans (2000): Deutsch in der Schweiz als nationale Varietät des Deutschen. In: Sprachreport, Heft 4, 21–27.

— (2006): Das Internet als linguistisches Korpus. In: Anton Näf & Rolf Duffner (Hg.), Korpuslinguistik im Zeitalter der Textdatenbanken, Linguistik online , 28, 71–83.

Bickel, Hans & Gasser, Markus & Häcki Buhofer, Annelies et al. (2009): Schweizer Text Korpus – Theoretische Grundlagen, Korpusdesign und Abfragemöglichkeiten. In: Annelies Häcki Buhofer (Hg.), Fortschritte in Sprach- und Textkorpusdesign und linguistischer Korpusanalyse II, Linguistik online, 39, 5–31.

Bickel, Hans & Schmidlin, Regula (2004): Ein Wörterbuch der nationalen und regionalen Varianten der deutschen Standardsprache. In: Thomas Studer & Günther Schneider (Hg.), Deutsch als Fremdsprache und Deutsch als Zweitsprache in der Schweiz, Bulletin vals-asla, 75, 99–122.

Perkuhn, Rainer & Belica, Cyril & al-Wadi, Doris et al. (2005): Korpustechnologie am Institut für Deutsche Sprache. In: Johannes Schwitalla & Werner Wegstein (Hg.): Korpuslinguistik deutsch: synchron - diachron - kontrastiv. Tübingen (Niemeyer).

Schmidlin, Regula (2003): Deutsch als plurizentrische Sprache: eine lexikographische und didaktische Herausforderung. In: Günther Schneider & Monika Clalüna (Hg.). Mehr Sprache – mehrsprachig – mit Deutsch. Didaktische und politische Perspektiven. München (Iudicium Verlag), 324-339

Wenger, Rosalia (1978). Rosalia G: Ein Leben. Bern (Zytglogge Verlag).