Korpusentwicklung und empirische Analysen zur Deskription eines Gebrauchsstandards

Elisabeth Scherr
Karl-Franzens-Universität Graz, Institut für Germanistik

Abstract. Das Projekt ‚Variantengrammatik des Standarddeutschen' ist ein internationales Forschungsunternehmen zur Dokumentation areal bedingter, grammatischer Variabilität im *Gebrauchsstandard* des Deutschen. Dieser Begriff von Standardsprachlichkeit nimmt die konkrete, in standardsprachlichen Kontexten eingesetzte Sprachgebrauchsnorm in den Fokus und bildet damit den deskriptiven Gegenpol zu normativen, strikt regelgeleiteten Orientierungen in Richtung einer homogenen Standardsprachlichkeit. Als Datengrundlage wird ein umfangreiches Korpus (285 Millionen Wortformen) erstellt, das sich vor allem durch die areale Ausgewogenheit des aufgenommenen Textmaterials auszeichnet, indem alle Voll- und Halbzentren des deutschen Sprachraums gleichwertig berücksichtigt werden. An deduktive und induktive Korpusrecherchen angeschlossen werden statistische Tests, um empirisch fundierte Aussagen zum Status der Variabilität im Gebrauchsstandard tätigen zu können. In der finalen Projektphase werden die Forschungsergebnisse in einem Handbuch zusammengefasst, das als variationslinguistische Grundlage für Forschung und Grammatikografie, aber auch als Nachschlagewerk für Laien konzipiert ist.

Keywords. Grammatik, Variation, Variantengrammatik, Varietäten, Varietätenlinguistik, Gebrauchsstandard, Pluriarealität, Korpuslinguistik

1. Projektkonzeption

Korpusbasierte Analysen und linguistische Deskriptionen zur areal bedingten, grammatischen Variation in standardsprachlichen Kontexten stellen grundsätzlich ein wissenschaftliches Desideratum dar. Während lexikalische Differenzen vergleichsweise gut bearbeitet sind (vgl. Ammon et al. 2004), soll im Rahmen des Projektes ‚Variantengrammatik des Standarddeutschen' diesem Bereich der Variation stärker Rechnung getragen werden (vgl. Dürscheid, Elspaß & Ziegler 2011). Als Ziel einer korpusgestützten Methodik werden Ergebnispräsentation und linguistische Beschreibung grammatischer, in standardsprachlichen Kontexten variabel in

[1] Das Projekt wird in Österreich gefördert vom Austrian Science Fund (FWF): Projektnummer I 716-G18.

Fuge in Nominalkomposita (z.B. bei *Rindsbraten* vs. *Rinderbraten*) in
Auflagen der Grammatiken zumindest erwähnt werden (ohne jedo
fundierte, differenzierte Aussagen tätigen zu können, vgl. bspw. Dud
Hentschel & Weydt 2003 zum Fugenmorphem; Zifonun et al. 1997 zu
Perfekt/Plusquamperfekt), bleibt eine Reihe von Auffälligkeiten auf de
Morphologie, der Morphosyntax oder der Syntax durchweg unb
folgenden Variantenbeispiele für den oberdeutschen Sprachraum (Beleg
CH-Zeitungstexten) sollen das weite Feld der grammatischen, standar
Variation in selektiver Auswahl illustrieren[2]:

- Wortbildung

(1) Letztlich soll der Neujahrsempfang aber dazu dienen, Kontakte zu pflegen und bei einem guten Glaserl Pläne für das Jahr zu schmieden. (meinbezirk v. 17.1.2011)

(2) Genießen Sie weiters schmackhafte Speisen und Getränke au der speziell kreierten Halloween Karte. (Bezirksblatt Hall in Tirol v. 25.10.2011)

(3) In Politikerkreisen wird jedoch angenommen, dass die EVP ihre Initiative zurückziehen wird, um den Gegenvorschlag nicht damit zu konkurrenzieren. (Zürichsee-Zeitung v. 5.10.2011)

- Flexion

(3) Freiwillig wohlgemerkt, zahlen gemusst hätte die Baugemeinschaft nicht – und das, obwohl der historische Kern ganz unmittelbar betroffen war. (Echo Tirol v. 1.10.2011

(4) Die Fussballer bekamen nicht einmal ein Rückreiseticket ausgestellt und lebten grösstenteils als illegale Immigranten i Rumänien. (20 Minuten v. 12.1.2012)

- Wort- und Satzgliedstellung

(5) Ist doch klar, dass Null-Kommunikation Konfliktpotenziale verstärkt und nicht verschwinden lässt. Möchte man meinen. (Echo Tirol v. 2.6.2009)

(6) Die aus diesbezüglich logischem Grund heile Dioxin-Welt Österreichs geriet wegen des Leutascher Kalbls gehörig ins Wanken. Für einen kurzen Augenblick zumindest. (Echo Tirc v. 1.2.2010)

Trotzdem, was ihm angetan wurde, trotzdem er kein reicher

[2] Sämtliche Beispiele stammen aus einem Pilotkorpus zur grammatischen Va Schweiz und in Österreich, das an der Universität Zürich erstellt wurde.

Diese Belege stellen nur einen kleinen Ausschnitt der grammatischen Teilgebiete dar, die areal bedingte Variation aufweisen. Auch in standardsprachlichen Kommunikationssituationen werden diese Varianten hoch frequent eingesetzt und sind daher in Modelltexten – im Rahmen des Projektes werden dies Pressetexte sein – nachzuweisen (Details zur Korpuskonzeption vgl. Punkt 3). In Beispiel (1) wird ein höchst produktives Muster zur Diminutivbildung präsentiert, das in Grammatiken meist bairischen Dialekten zugeordnet wird (vgl. Hentschel & Weydt 2003: 197). Der Zweifelsfälle-Duden führt als standardsprachlich nur die Diminutivvarianten *–chen* und *–lein* an und verweist auf die „emotionale Beteiligung" (2011: 240) des Sprechers. Eine erste Korpusrecherche zeigt jedoch, dass das Diminutivbildungsmuster mit *–erl* für den oberdeutschen Sprachraum auch in standardsprachlichen Kontexten höchst frequent ist (u.a. *Stamperl, Schnapserl, Zuckerl, Stüberl*) und nicht nur in markierten Kontexten (d.h. nicht nur bei emotionaler Beteiligung oder in direkten Zitaten) verwendet wird. Beispiel (2) zeigt ein Wortbildungsmuster der Adverbien, das in Grammatiken oft mit dem Label „umgangssprachlich"/„regional" oder „gesprochensprachlich" versehen wird (vgl. Wahrig 2006: 289; Zweifelsfälle-Duden 2011: 1012). Beispiel (3) ist ein Beleg zur regelmäßigen Partizipialbildung des Modalverbs, an dieser Stelle müsste nach normativer Grammatik der Ersatzinfinitiv stehen, so gefordert auch bei Engel (1996: 464): „Bei den Modalverben hört man in der Alltagssprache gelegentlich regelmäßig gebildete Partizipien; sie gelten jedoch standardsprachlich als unkorrekt […]." Der Ausbau des Passivparadigmas durch eine Konstruktion des Rezipientenpassivs wird durch Beispiel (4) illustriert. Das Verb *bekommen* (ähnlich den Verben *erhalten* oder *kriegen* in analogen Konstruktionen) nimmt hier Auxiliarbedeutung an und dient der Passivierung eine Dativobjektes (meist in der semantischen Rolle des Benefaktivs), der in Subjektsposition erscheint (vgl. Wahrig 2006: 67). Satz (5) zeigt die Extrapositionierung eines Teilsatzes ohne Vorfeldbesetzung, ein Schema, das vor allem in Schweizer Zeitungstexten auffällig oft zu belegen ist. Auch der angeschlossene Satz desselben Beispiels zeigt diese Variante. Es handelt sich hierbei um einen elliptischen Nachsatz (wie auch in Beispiel (6) zu sehen ist), bei dem es zu kontextuell bedingter Aussparung meist eines phorischen Elements oder der gesamten Verbalphrase kommt. In Beispiel (7) findet eine Abweichung von der normativen Konstituentenabfolge im Verbalkomplex statt. Wiewohl diese Tatsache bereits vermehrt behandelt wurde, geschah dies in den meisten Fällen im Zuge der Dialektforschung (vgl. Patocka 1997: 279f), Abweichungen sind jedoch auch in standardsprachlichen Kontexten höchst frequent belegt. Beispiel (8) schließlich zeigt die Verwendung eines Interrogativpronomens als Indefinitpronomen. Dies betrifft auch die Pronomina *wer* oder *eine*, eine Verwendungsweise, die von Grammatiken als umgangssprachlich klassifiziert wird: „Umgangssprachlich können auch die Interrogativa *wer* und *welch* als

Das Projekt ‚Variantengrammatik des Standarddeutschen' ist ein
drei Jahre angelegtes, trinationales Kooperationsunternehmen
Projektstandorten Universität Zürich, Universität Graz und Universität S
dieser Konzeption ergibt sich auch die Dreiteilung der Projektleitung
durch Prof. Dr. Arne Ziegler, in Zürich durch Prof. Dr. Christa Dürs
Salzburg durch Prof. Dr. Stephan Elspaß repräsentiert ist. Zusätzlich
jeden Projektstandort zwei Doktorandenstellen geschaffen, um
weiterführende Dissertationen im Umfeld der Variantengrammatik r
können. Finanziert wird dieses internationale Vorhaben durch die
dingungen des D-A-CH-Abkommens, das zwischen den Förderins
Deutschland (Deutsche Forschungsgemeinschaft DFG), Österrei
schaftsfonds FWF) und der Schweiz (Schweizer Nationalfonds SNF) best

Dem Projekt liegt als theoretische Fundierung der Begriff des „G
dards"(Ammon 1995: 88) zugrunde, wie er unter Punkt 2 dargelegt wer
der sich durch die Anerkennung eines Variationspotenzials auch i
Standardsprachlichkeit auszeichnet. In diesem Sinne wird ein Ansatz
sich einerseits dezidiert vom Ideal einer variationsfreien, strikt nor
Standardsprache abgrenzt, andererseits jedoch auch ideologischen, s
orientierten Konzepten entgegensteht.

2. Eine Frage des Standards

Das Konzept einer homogenen ‚Standardsprache' wird auch innerhalb d
linguistik oft als einheitlicher Referenz- oder Ausgangspunkt für di
weniger stark sanktionierten Abweichungen der Varietäten herang
Dürscheid, Elspaß & Ziegler 2011: 123). Es liegt jedoch auf der Hand,
oben dargestellten Blickwinkel ein anderer Standardbegriff zugrunde g
muss, durch den der arealen Variabilität stärker Rechnung getrage
Aufgabe von ‚Standard' im präskriptiven Sinne zugunsten eines offene
Standardvarietät, die mehr ist als „das, was uns die kodifizierte präsk
glauben machen will" (Ziegler 2011: 251), führt zu einer Interpreta
tatsächlichen Sprachgebrauch stärker berücksichtigt. Die Variabilität vo
heterogenes aber geordnetes System betrifft dabei alle Re
Variationsebenen, die standardsprachlichen Varietäten bilden dabei kein
Ihr spezifisches Charakteristikum ist allerdings durch das entscheid
geprägt, dass es sich bei den Variationsausprägungen nicht um Merkmal
oder diaphasisch beeinflusster *Non*standardvarietäten handelt, die
einheitlichen Standardsprache im weitesten Sinne „überdacht" (Amm
werden. Vielmehr geht es um die empirisch-deskriptive Sich

Variationsmuster dabei keineswegs mit nationalen Landesgrenzen koinzidieren und noch dazu selten innerhalb politischer Grenzen in homogenem Gebrauch sind, muss die Bezeichnung Pluri*nationalität* als inadäquat gesehen werden. Angemessener erscheint der Terminus „Pluriarealität" (Wolf 1994), der keine Gebundenheit der Varianten an Staatsgrenzen suggeriert (vgl. Dürscheid, Elspaß & Ziegler 2011). Diese sprachgebrauchsbasierten Ausprägungen der heterogenen Varianten der Standardsprachlichkeit, deren Grenzen fließend ineinandergreifen, sind dabei charakterisiert durch die Tatsache, dass sie nicht unbedingt durch den Sprachkodex als standardsprachlich ausgewiesen sind, sondern dass zumindest Modelltexte zugrunde liegen, die deren Gebrauch in hinreichender (d.h. signifikanter) Weise belegen. Für die Bildung einer Basis zur Analyse eines Gebrauchsstandards wurden dabei Pressetexte ausgewählt, denen Modellcharakter im Hinblick auf Standardsprachlichkeit zugeschrieben wird (vgl. Ammon 2005: 88). Mit Eisenberg (2007: 217) wird davon ausgegangen, dass der Status einer Variante „statistisch durch eine Auswertung umfangreicher Zeitungskorpora zu ermitteln ist". Ohne die normative Kraft von Standardsprachlichkeit in Frage stellen zu wollen, dienen die beiden – ebenfalls von Eisenberg formulierten – Prämissen als Ausgangspunkt für die Korpuserstellung und als methodisches Gerüst für das Projektvorhaben:

> Der erste [Schritt] besteht darin, dem geschriebenen Standard als normsetzender Leitvarietät Geltung zu verschaffen. Der zweite besteht in der Übereinkunft (manchmal möchte man sagen: in dem Zugeständnis), dass der Sprachgebrauch innerhalb dieser Varietät empirisch erhoben werden kann und dass deshalb vorwissenschaftliche Gewissheiten nicht anerkannt werden. (Eisenberg 2009: 64)

3. Korpuskonzeption

Zur systematischen Dokumentation der grammatischen Variation wird daher im Rahmen des Forschungsprojektes ein Großkorpus aus Zeitungstexten des gesamten deutschen Sprachgebietes erstellt, das als empirische Grundlage dient. Die mangelnde areale Ausgewogenheit und die größenmäßige Beschränkung bereits bestehender Korpora (DeReKo mit Cosmas II oder Korpus c4) sowie die oft eingeschränkten Nutzungsrechte führten zum projektinternen Beschluss, ein eigenes, für variantenlinguistische Zwecke gut nutzbares Großkorpus aufzubauen.[3] Zur Textauswahl wurde das gesamte deutsche Sprachgebiet (unter Berücksichtigung

[3] Allerdings werden die o.g. Korpora als Referenzkorpora genutzt werden, z.B. um typische Verwendungsweisen in zweiter Instanz überprüfen oder Kontrollabfragen zum tentativen Vergleich der Rechercheergebnisse durchführen zu können.

werden kann, dass die Variantengrenzen mit den definierten area
korrespondieren.

Im Grunde wurde die Einteilung des deutschsprachigen Rau
Erstauflage des Variantenwörterbuchs übernommen (vgl. Ammon et al.
ff.), in dem Deutschland in sechs und Österreich in vier Sektoren einget
Liechtenstein, Luxemburg, Ost-Belgien und Südtirol als eigene Geb
werden. Erweitert wurde diese Systematisierung jedoch durch eine D
Schweiz in einen nordwestlichen, einen nordöstlichen und in einen südl
neuere Untersuchungen zur Binnendifferenzierung der dialektalen S
Schweiz (vgl. Seiler 2005) eine solche Untergliederung auch für die
Standardsprache als lohnend erscheinen lassen.

Abb. 1: Die projektinterne Einteilung der Schweiz in Anlehnung an
Dialektgliederung (vgl. Glaser 2003; Seiler 2005)

Auch aus Gründen der Praktikabilität wird auf die Onlineausgaben der
überregionaler Zeitungen zurückgegriffen, die Berücksicht
Verbreitungsgebiete der jeweiligen Printversionen garantiert dabei zu
gleichmäßige geografische Erfassung. Dieses Vorgehen rechtfertigt
mehrere Gründe: Durch die Klassifikation der Zeitungen als *überregi*
Standardsprachlichkeit des Textmaterials sichergestellt. Auf der andere
von unmittelbarer Relevanz, dass die Autoren der betreffenden
tatsächlich aus der Region stammen, für die sie schreiben. Di
überregionalen Gesamtteils bestehen oft aus Meldungen der Presseag
werden in zentralen Redaktionen verfasst. Nur ein Ausschluss dieser Ar
Großkorpus ermöglicht es, Aussagen bezüglich der arealen Variabilität
Bewusstsein der Wichtigkeit dieses Punktes wurde in der konzeptionel

Pro Sektor wird die exakt gleiche Textmenge aus den Online-Ausgaben der Regionalteile der überregionalen Tageszeitungen zunächst in eine Rohdatenbank eingespeist. Insgesamt wurden 57 Zeitungen – bzw. deren Regionalteile – ausgewählt und mittels URL-Erkennung in den Crawler eingebaut. Bei einer Textmenge von fünf Millionen Wortformen pro Zeitung ergibt das eine Gesamtkorpusgröße von 285 Millionen Tokens. Zur Vermeidung textsortengebundener Variabilität wurden einerseits verschiedene Textsorten mit verschiedenen Formalitätsgraden berücksichtigt. Ausgeblendet blieben andererseits Werbetexte, Anzeigen und Agenturmeldungen, da deren sprachliche Beschaffenheit als zu spezifisch bzw. als inadäquat für eine variantenlinguistische Untersuchung zur Gebrauchsstandard-sprachlichkeit eingestuft wurde (vgl. Dürscheid, Elspaß & Ziegler 2011: 135). Durch das korpuslinguistische Methodenbewusstsein der quantitativen Linguistik wurde versucht, die Problematiken der empirischen Wissenschaften so gut es geht einzudämmen, die sich vor allem in den folgenden Punkten manifestieren (vgl. Köhler 2005a; 2005b):

1. *Repräsentativität*: Kein Korpus kann groß genug sein, um eine wie auch immer geartete ‚Sprachwirklichkeit' abzubilden. Durch die Aufnahme möglichst verschiedenartiger Texte (Reportagen, Berichte, Dossiers, Nachrichtenmeldungen) wurde versucht, die Stilebenen möglichst breit zu halten, um textsortenspezifische Varianz auszugleichen. Andererseits findet dezidiert eine Beschränkung auf den *geschriebenen* Standard statt, was Textsorten wie Glossen, Interviews oder Kommentare ausschließt.
2. *Homogenität* und *Homöoskedastizität* (gleichbleibende Varianz): Auch dieses Kriterium ist für Sprachdaten meist nicht bis ungenügend erfüllt. Nur selten sind sprachliche Phänomene normalverteilt, was sich vor allem in der Beschreibung von grammatischen Varianten manifestiert. Es kann nicht davon ausgegangen werden, dass es zu einer homogen Verteilung der Daten kommt, oder dass der Grad an Varianz in allen Bereichen gleich groß ist. Statistische Tests werden im Rahmen des Forschungsprojektes auch nicht dazu verwendet werden, Gleichmäßigkeit zu beschreiben, sondern im Gegenteil dazu, Variabilität und Schwankungsbreiten auch in berechenbare Zahlen zu fassen. (Zu statistischen Methoden vgl. Punkt 6)
3. *Extreme Schiefe*: Dieses für Sprache sehr typische Phänomen besteht darin, dass seltene Einheiten in Textsammlungen stets unterrepräsentiert sind. Durch die umfassende Größe des Korpus wurde versucht, auch diesem Problem Rechnung zu tragen. Die lemmaunabhängige Recherche wird zusätzlich eine Möglichkeit bieten, ein größeres Spektrum von Konstruktionsweisen zu erfassen (vgl. Punkt 5 zur Analyse).

der Forschergruppe *semtracks* die Korpuserstellung durchzuführen. Mitglied der *Black Forest Grid* der Universität Freiburg, eine interdiszipl von Forschern, die sich mit computerbasierten Analysen in de auseinandersetzt (vgl. Homepage Universität Freiburg: http://v freiburg.de/). Die Zusammenstellung des Korpus mittels Parallelisierun es, sehr große Textmengen in relativ kurzer Zeit zu archivieren, zu an abzufragen, dies gelingt mit klassischer, sequentieller (od Programmierung nur begrenzt für das Ausmaß der Datenmenge des *semtracks* hat zudem Erfahrung darin, Korpora über ein Benutzeroberfläche (GUI) zugänglich zu machen. Durch die digitale der Zeitungsberichte müssen die Texte nur von einem Crawler erfasst temporären Internetspeicher in eine Datenbank übertragen werden. Als *.arc*-Dateien (Archiv-Dateien) werden sie in eine zentrale Korp überführt und mit der Suchsoftware *Lucene* ausgestattet, bevor die Annotation durchgeführt werden kann. Im Endergebnis stehen für mindestens fünf Annotationstypen zur Verfügung, die für jedes To wurden und in Verwendung der CQL (*Corpus Query Language*) na abgefragt werden können:

- **[word=]**: Mit diesem Syntaxbefehl kann eine Wortform gesucht zwar so, wie sie im Text erscheint. Dies dient der wortformgel und der Feststellung von absoluten Häufigkeiten. Mit dem [word="Häuser"] werden alle Vorkommen der in Anfüh spezifizierten Wortform gesucht. Während dieses Annotations Feststellung von Einzelfrequenzen verwendet werden kan systematische, phänomenorientierte Suche damit nur s durchzuführen, da die definierte Wortform sehr eingeschränk liefert.
- **[pos=]**: Die Suche nach positionellen Attributen liefert ein werkzeug namens TreeTagger, das mit diesem Befehl aktiv werden sehr simple Wortartinformationen angezeigt, die zusammengefasst und durch das Stuttgart-Tübingen-Tagset wurden, hier ein paar Beispiele in Auswahl:

ADJA	attributives Adjektiv
ADJD	adverbiales oder prädikatives Adjektiv
ART	Artikel
CARD	Kardinalzahl
NN	normales Nomen
NE	Eigennamen
PTKZU	"zu" vor Infinitiv
VVFIN	finites Verb, voll

VMPP Partizip Perfekt, modal
(vgl. STTS-Tagset des IMS Stuttgart: http://www.ims.uni-stuttgart.de/projekte/corplex/TagSets/stts-table.html)

Diese Attribute können mit dem Suchbefehl [pos=] gesucht werden, wobei es zu keiner detaillierteren Charakterisierung eines Token kommen kann, das heißt Tempus- oder Genusinformationen können nicht separat spezifiziert werden. Um die Suchergebnisse stärker einzuschränken, bietet das folgende Annotationsinstrument differenziertere Möglichkeiten.

- **[rfpos=]**: Der so genannte RF-Tagger liefert morphosyntaktische Annotationen, die es erlauben, komplexe Suchanfragen zu starten, indem mehrere verschiedene Attribute für eine Wortform festgelegt werden können:

Abb. 2: Die unterschiedlichen Wertzuweisungen des RF-Taggers

Merkmal	Wertebereich
KOMPARATION	{Comp. Pos. Sup}
KASUS	{Nom. Gen. Dat. Acc. *}
NUMERUS	{Sg. Pl. *}
GENUS	{Fem. Masc. Neut. *}
PERSON	{1. 2. 3. -}
TEMPUS	{Past. Pres}
DETERMINATION	{Def. Indef}
PREP	{Nom. Gen. Dat. Acc. Als. An. Außer. Auf. Bis. Hinter. In. Je. Namens. Per. Pro. Unter. Vor. -. ?}
CONJ	{Comp. Coord. SubFin. SubInf}
CONJCOORD	{Aber. Als. Bis. Denn. Doch. Noch. Wie}
PART	{Ans. Deg. Neg. Verb. Zu}
PRO	{Dem. Indef. Pers. Refl. Rel}
PROADV	{Dem. Inter}
SUBST	{Subst. Attr}
SYM	{Other. Paren. Pun. Quot}
SYMSPEC	{Aster. Auth. XY. Left. Right. Colon. Comma. Cont. Hyph. Sent. Slash}
PARTES	{Adj. Noun. Verb}
VERB	{Aux. Full. Haben. Mod. Sein}
ZU	{zu. -}
PARTICIPIUM	{Psp}

So wird es ermöglicht, für jedes Merkmal detailliertere Attribute zu spezifizieren, was die Suchanfrage zwar komplexer gestaltet, dafür jedoch eine überschaubare Menge an Ergebnissen liefert.

- **[lemma=]**: Für jedes Token stehen mit diesem Suchbefehl Lemmainformationen zur Verfügung, das ist vor allem für Phänomene relevant, die eine lemmaorientierte Suche erfordern, beispielsweise die Analyse sämtlicher Flexionsformen.
- **[morph=]**: Die morphologische Analyse, die Abfragen zur Wortbildung erlaubt, wird von einem Programm namens „Morphisto“ durchgeführt. So können Wortbildungsphänomene wie Fugenrealisierungen oder Derivationsmuster gesucht werden.

5. Analysemethoden: *corpusdriven/corpus*

Grundsätzlich wird die Korpusarbeit in zwei Schritten vor sich ge
insgesamt eine Kombination aus induktivem (*corpus-driven*) und
(*corpus-based*) Zugang angestrebt (vgl. Bubenhofer 2006). Der de
korpusbasierte Arbeitsschritt besteht darin, das Korpus nach v
Suchmustern zu untersuchen, das heißt nach Auffälligkeiten, d
projektinternen Auflistung dokumentiert wurden; diese Liste umfasst na
Stand zirka 750 Varianten. Die Erarbeitung dieser Auflistung gesch
systematische Überprüfung vorliegender grammatischer Nachschla
deutschen Gegenwartssprache sowie einschlägiger Arbeiten aus dem
Variantenlinguistik. In diesem Arbeitsschritt fand eine Durchsicht der
eine Dokumentation von Verweisen wie „dialektal" „südde
„umgangssprachlich", aber auch „österreichisch" oder „schweizerisch"
sich dann bei dem betreffenden Phänomen auch tatsächlich um eine
Gebrauchsstandards handelt, muss natürlich noch unbeantwortet bleib
soll die korpusbasierte Analyse in Kombination mit statistischen A
zeigen (vgl. Punkt 6). Diese Sammlung auffälliger Abweichungen wird
durch die Durchsicht noch fehlender, relevanter Werke und Neuaufl
erweitert, andererseits aber auch durch Varianten ergänzt, die bei
Schweizer und österreichischer Zeitungen augenfällig werden. Diese A
von Mitarbeitern des Projektstandortes Graz für die Schweizer Texte v
in Zürich werden österreichische Zeitungen gelesen, um Auffälligkeiten
des anderen Sprachgebietes besser beurteilen zu können. Diese Doku
werden dann systematisiert und ebenfalls in die Gesamtliste einge
Prüfung auf Signifikanz erfolgt dann wiederum durch die
‚mutmaßlichen' Varianten am Großkorpus. Dieses Verfahren der
(korpusbasierten) Analyse dient dazu, eine Hypothese, beispielswei
„Auffälligkeit xy ist eine gebrauchsstandardsprachliche Variante"
Korpusdaten zu stützen. Wiewohl durch dieses Verfahren eindeuti
Belege für introspektive Vermutungen zu belegen sind, hat es natürlic
entscheidenden Nachteil: Die deskriptiven Bewertungen verlassen nie
mutmaßlichen Varianten, die bereits an anderer Stelle thematisiert wu
salient sind, dass sie beim Lesen von Zeitungstexten auffallen. D
Korpusrecherche besteht dann darin, zwar die Liste von Varianten
trotzdem bleiben systematische (aber weniger saliente) Abweichungen u

Aufgrund dieser Tatsache werden die Recherchen noch zusätzlich u
der *corpus-driven* Analyse erweitert, und zwar durch die Anwendung
‚n-Gramm'-Berechnungen, die sich durch eine Kombination
kationsanalysen und der syntagmatischen Muster auszeichnen und a

Herangehensweise ist insofern attraktiv, als dass sie Hinweise auf Varianten oder Variantenpräferenzen geben könnte, die bisher nicht im Fokus lagen. Dies wird dadurch ermöglicht, dass gewisse Auffälligkeiten im Korpus automatisiert angezeigt werden können, ohne strikt lemmabasiert vorgehen zu müssen. Ein Beispiel soll dieses Vorgehen illustrieren: Im Hinblick auf die Verwendung des Reflexivums im Deutschen sind folgende Konstruktionsvarianten augenscheinlich, die verstärkt im oberdeutschen Gebiet verwendet werden:

- Nach Wochen beschließt er, **sich eine Hütte zu bauen**, eine Höhle wählt er als Ausweichquartier. (Echo Tirol v. 1.2.2009)
- Gerade beim Geld **hört sich der Spass** bei den meisten Menschen **auf**. (meinbezirk v. 25.11.2011)

Da die semantische Rolle des Reflexivums in diesen Fällen nicht eindeutig bestimmt ist, kann nicht von einer Valenzerweiterung des Verbs ausgegangen werden, vielmehr könnte es sich um eine polyfunktionale Verwendung des *sich* als Medialmarker handeln, indem es zur stärkeren Subjekteinbindung in das Verbalgeschehen kommt. Eine tiefergehende Analyse des Phänomens ist an dieser Stelle zweitrangig[4], vielmehr soll nach einer empirisch validen Recherchemöglichkeit zur Deskription des Phänomens gesucht werden. Nun kann in korpuslinguistischer Herangehensweise natürlich eine Suche nach den Kombinationen der Verben (*aufhören, bauen*) mit dem Reflexivum erfolgen, um zu beschreiben, in welcher Verwendungsweise es eingesetzt wird. Dabei wird jedoch der Blick verbaut auf eventuell andere Verben, die ebenso in dieser Konstruktion vorkommen. Aus diesem Grund wird eine Reduktion der Einheiten auf deren Attributwerte vorgenommen, um so nach den spezifischen Merkmalsbündeln – Kookurrenzen der Annotationen, wenn man so will – zu suchen. In sehr vereinfachender Darstellung kann das folgendermaßen veranschaulicht werden:

Der Trainer erwartet sich vollen Einsatz.
der Trainer / erwarten / sich / voll / Einsatz
der Trainer / erwarten / sich / voll / NN
der Trainer / erwarten / sich / ADJA / NN
der Trainer / VV / PRF / ADJA / NN
NN / VV / PRF / ADJA / NN
(vgl. Bubenhofer 2009)

Ausgehend also von der Erkenntnis, dass das Reflexivum in dieser speziellen zu untersuchenden Konstruktionsweise vorkommt, werden „statistisch signifikante sekundäre Kookkurrenzen berechnet“ (Bubenhofer 2009: 119) und zu syntag-

[4] Zur systematischen Beschreibung dieser Tendenz vgl. Àgel 1997, 2000; Ziegler 2010.

starten, die Ergebnisse ohne eindeutige Attributwertzuschreibung
Ergebnisse, bei denen sich die Tools ‚unsicher' in der Charakterisierun
wird auf die Tatsache zurückgeführt, dass das Suchsystem mit vorgefe
arbeitet, die beispielsweise sämtliche Präteritumsformen der deuts
beinhalten, die natürlich nur aus den Standardflexionen bestehen. Dad
Möglichkeit eröffnet, sämtliche Präteritumsformen im Korpus abzuruf
eindeutige Zuschreibung haben, und sie entweder automatisiert mit e
Standardformen vergleichen zu lassen, oder die ambigen Fälle direkt
Suchbefehl auszugeben. Dieses Vorgehen bringt nun ebenfalls völlig ne
die bisher noch nicht behandelt wurden. So wird eine umfassende De
Variationsspektrums ermöglicht, die von der Forschung bisher in die
noch nicht vorgenommen wurde.

6. Zur Signifikanz variabler Fallzahlen

Ein Problem bei der oben dargestellten Form der deskriptiven Korpusa
jedoch: Die Ermittlung der Varianten und deren Frequenz im Korpus lie
Suchanfrage nur absolute Frequenzen, dadurch ist noch relativ we
Frequenzverteilung, geschweige denn über die Signifikanz der Variant
Ein einfaches Beispiel soll das Problem illustrieren: Es gibt im Bereich
Wortbildung eine Reihe von Auffälligkeiten, die vor allem der
oberdeutschen Raum betreffen. Verben der Form *parkieren* sind in der S
unmarkiert, während weiter im Osten oder im norddeutschen Sp
standardsprachlichen Kontexten *parken* verwendet wird. Durch die A
Textmaterials ist es nun vergleichsweise einfach, zwei Suchanfragen zu
häufigkeit zu starten. In dem bereits oben charakterisierten Probekorp
nun folgende Ergebnisse:

Tab. 1: Suchergebnisse für *parken/parkieren*

	CH [in 8,753,571 Wortformen]		**D** [11,510,491 Wortforme	
	parkieren	*parken*	*parkieren*	*parken*
Frequenz	46	14	33	76

Nun sagen diese Zahlen natürlich nicht viel aus, da die Wortformen
ungleichmäßig auf die beiden Länder verteilt sind. Die Berechnung
Häufigkeit pro Subkorpus ergibt für eine Million Wortformen folgende W

Tab. 2: Relative Häufigkeiten für *parken/parkieren*

	CH [pro 1 Mio. Wortformen]		**D** [pro 1 Mio. Wortfo	
	parkieren	*parken*	*parkieren*	*parke*
rel. Häufigkeit	5,25	1,6	2,86	6,6

Tritt *parkieren* in der Schweiz signifikant häufig auf

- im Vergleich zu *parken* in der Schweiz und/oder
- im Vergleich zu *parkieren* in Deutschland?

Damit ist auch eine grundlegende Fragestellung von quantitativen Analysen angesprochen, die zum Testen von Hypothesen in den meisten Fällen beantwortet werden soll, nämlich die Frage, ob eine Korrelation zwischen zwei Variablen festgestellt und als statistisch signifikant eingestuft werden kann (vgl. Bubenhofer 2009: 134). Der statistische Loglikelihood-Test, der t-Test oder der Chi-Quadrat-Test (vgl. Gries 2008: 187ff) können im Grunde darüber Aufschluss geben, ob eine Variante in einer so hohen Frequenz auftritt, dass die Abweichung von einer zufälligen Verteilung signifikant ist. Dazu werden Kontingenztabellen für die beobachteten und die erwarteten Werte erstellt, um auf deren Basis den Signifikanzwert unter Berücksichtigung des Freiheitsgrades zu errechnen. Dabei ergeben sich jedoch im Falle der sprachlichen Korpusdaten entscheidende Problematiken, die auf die Spezifika einer textbasierten Analyse zurückzuführen sind: Beim Untersuchungsgegenstand ‚Sprache' kann nie davon ausgegangen werden, dass eine zufällige Verteilung von Einzelelementen vorliegt, was ein Verwerfen der Nullhypothese in nahezu allen Fällen nach sich zieht. Entweder muss also ein statistischer Text gewählt werden, der dem Problem der „Klumpen" (Bubenhofer 2009: 134) der sprachlichen Entitäten mit der Teilung des Gesamtkorpus in Untergruppierungen begegnet, dies leistet beispielsweise der Mann-Whitney-Test (vgl. Kilgarriff 2001: 103). Eine andere Möglichkeit besteht in der methodologischen Herangehensweise, die oben genannten statistischen Signifikanztest lediglich dazu zu verwenden, den Grad der Unzufälligkeit eines bestimmten Vorkommens herzustellen, ohne automatisch auf Signifikanz zu schließen (vgl. Bubenhofer 2009: 135). So gelingt es, zumindest eine Rangordnung der Varianten herzustellen, die Wahl eines Tests muss dabei phänomenorientiert erfolgen, da bei unterschiedlichen Fallzahlen mit unterschiedlichen Konstruktionsweisen der grammatischen Varianten verschiedene Verfahren angewendet werden sollten, um verlässliche Aussagen zu erhalten.

Grundsätzlich kann davon ausgegangen werden, dass die Korpusgröße ein entscheidendes Kriterium bleibt, um Entscheidungen betreffend der standardsprachlichen Signifikanz einer Variante treffen zu können. Durch die umfassende Erhebung von Datenmengen wird es ermöglicht, zwischen Motiviertheit und Arbitrarität von Einzelphänomenen wenigstens tendenziell entscheiden zu können, ohne auf einen umfassenden Ausgleich durch statistische Verfahren angewiesen zu sein. Unterstützend können Methoden zur Feststellung der Relevanz oder zur Erhebung von Kollokationen als Ergänzungen zum rein deduktiven Vorgehen angeschlossen werden.

Mangel durch die Anwendung von deskriptiven, empirischen Method
Die Dokumentation grammatischer Varianten in einem Handbuch, das
Arbeitsphase des Projektes entstehen soll, leistet langfristig ein Beitrag
regionalen Ausprägungen des Standarddeutschen als gleichwertig anerk
In dieser Hinsicht hat das Projekt eine zentrale sprachpolitische Bedeutu
die Grammatikografie des Deutschen profitiert von der Unternehmung.
bis jetzt noch auf keine einzige umfassende empirische Untersuchung
in der Standardgrammatik stützen (vgl. Dürscheid, Elspaß & Ziegler 20
geplante Handbuch wird dabei einerseits nach objektsprachliche
(konkreten Einzelfällen) und andererseits nach grammatischen
(übergeordneten Themen) gegliedert sein. So wird die Möglichkeit g
unter dem Stichwort *(sich) erwarten* die Angabe der Konstruktion als ty
oberdeutschen Sprachraum – vielleicht aber nicht generell – nach
entsprechende Übersichtsartikel zum Stichwort *Reflexivität* gi
Informationen zur auffälligen Präferenz in differenten Arealen d
Sprachgebietes.

Durch die Erarbeitung eines areal ausgewogenen, umfangreiche
wird auch für zukünftige Untersuchungen zur Variation in d
Standardsprache eine Grundlage geschaffen, die nach Projektende ei
Öffentlichkeit zugänglich gemacht werden soll. Durch die d
Annotationsmethoden dient das Datenmaterial für die Grammatikfo
gleichzeitig für deskriptive, computerbasierte Korpusanalysen
Ausgangsbasis. Die Beschreibungen zu variablen Sprachphänomenen, s
auf eine empirisch fundierte Basis zurückgeführt, werden all jenen Nut
sein, die sich aus wissenschaftlichen und/oder rein interessensgeleiteten
Variation innerhalb des deutschen Gebrauchsstandards beschäftigen wo

Weitere projektrelevante Informationen unter:
http://www.variantengrammatik.net
http://www.semtracks.org/web/
http://www.uni-graz.at/stage/germahww/germahww_aktuelles.htm

8. Literatur

Ágel, Vilmos (1997). Reflexiv-Passiv, das (im Deutschen) keines ist. Übe
Reflexivität, Medialität, Passiv und Subjekt. In Christa Dürscheid
Ramers & Monika Schwarz (Hgg.), *Sprache im Fokus*. Tübingen: N
187.

Ágel, Vilmos (2000). *Valenztheorie*. Tübingen: Narr.

Deutschen. Die deutsche Standardsprache in Österreich, der Schweiz und Deutschland sowie in Liechtenstein, Luxemburg, Ostbelgien und Südtirol. Berlin/New York: de Gruyter.

Bubenhofer, Noah (2006). *Einführung in die Korpuslinguistik. Praktische Grundlage und Werkzeuge.* Online im Internet: URL: http://www.bubenhofer.com/korpuslinguistik/kurs/index.php?id=uebersicht.html [2011-12-20]

Bubenhofer, Noah (2009). *Sprachgebrauchsmuster. Korpuslinguistik als Methode der Diskurs- und Kulturanalyse.* Berlin/New York: de Gruyter.

Duden. Die Grammatik. Unentbehrlich für richtiges Deutsch. 8., überarbeitete Auflage 2009. Mannheim/Wien/Zürich: Dudenverlag.

Duden. Richtiges und gutes Deutsch. Das Wörterbuch der sprachlichen Zweifelsfälle. 7., vollständig überarbeitete Auflage 2011. Mannheim/Wien/Zürich: Dudenverlag.

Dürscheid, Christa, Stephan Elspaß & Arne Ziegler (2011). Grammatische Variabilität im Gebrauchsstandard: das Projekt ‚Variantengrammatik des Standarddeutschen'. In Marek Konopka, Jacqueline Kubczak, Christian Mair, František Štícha & Ulrich H. Waßner (Hgg.), *Grammatik und Korpora 2009 / Grammar & Corpora 2009.* Tübingen: Narr, pp. 123-140.

Eisenberg, Peter (2007). Sprachliches Wissen im *Wörterbuch der Zweifelsfälle.* Über die Rekonstruktion einer Gebrauchsnorm. *Aptum,* 3, 209-228.

Eisenberg, Peter (2009). Richtig gutes und richtig schlechtes Deutsch. In Marek Konopka & Bruno Strecker (Hgg.), *Deutsche Grammatik. Regeln, Normen, Sprachgebrauch.* Berlin/New York: de Gruyter, pp. 53-69.

Engel, Ulrich (1996). *Deutsche Grammatik.* 3., korrigierte Auflage. Heidelberg: Groos.

Glaser, Elvira (2003). Schweizerdeutsche Syntax. Phänomene und Entwicklungen. In Beat Dittli et al. (Hgg.), *Gömmer MiGro? Veränderungen und Entwicklungen im heutigen SchweizerDeutschen.* Freiburg: Universitätsverlag Freiburg, pp. 39-66.

Gries, Stefan (2009). *Statistik für Sprachwissenschaftler.* Göttingen: Vandenhoeck & Ruprecht.

Helbig, Gerhard & Joachim Buscha (2001). *Deutsche Grammatik. Ein Handbuch für den Ausländerunterricht.* Berlin/München: Langenscheidt.

Hentschel, Elke & Harald Weydt (2003). *Handbuch der deutschen Grammatik.* 3., völlig neu bearbeitete Auflage. Berlin/New York: de Gruyter.

Kilgarriff, Adam (2001). Comparing Corpora. *International Journal of Corpus Linguistics*, 6, 1-37.

Köhler, Reinhard (2005). Gegenstand und Arbeitsweise der Quantitativen Linguistik. In Reinhard Köhler, Gabriel Altman & Rajmund G. Piotrowski (Hgg.), *Quantitative Linguistik. Ein internationales Handbuch.* Berlin/New York: de Gruyter.

Frankfurt am Main: Lang.
Seiler, Guido (2005). Wie verlaufen syntaktische Isoglossen, Konsequenzen sind daraus zu ziehen? In Eckhard Eggers et al. (Hg *Dialekte – Neue Dialektologie*. Stuttgart: Steiner, pp. 313-341.
Wahrig. Die deutsche Rechtschreibung (2006). Hrsg. von der Wahr Wissenschaftliche Beratung durch Lutz Götze. Gütersloh /Münche Verlag.
Wolf, Norbert Richard (1994). Österreichisches zum österreichischen Anlaß des Erscheinens von Wolfgang Pollack: Was halten die Öst ihrem Deutsch? *Zeitschrift für Dialektologie und Linguistik,* 61, 66-7
Ziegler, Arne (2010). ‚Er erwartet sich nur das Beste' … Reflexivierung Ausbau des Verbalparadigmas in der österreichischen Standar Dagmar Bittner & Livio Gaeta (Hgg.), *Kodierungstechniken im Zusammenspiel von Analytik und Synthese im Gegenwartsdeutsche* York: de Gruyter, pp. 65-81.
Ziegler, Arne (2011). Standardsprachliche Variation und grammatische Klaus-Michael Köpcke & Arne Ziegler (Hgg.), *Grammatik – Le Verstehen. Zugänge zur Grammatik des Gegenwartsdeutschen.* Berl de Gruyter, pp. 245-264.
Zifonun, Gisela, Ludger Hoffmann & Bruno Strecker u.a. (1997). *Gr deutschen Sprache*. Berlin/New York: de Gruyter.