V: auf Anfrage

Züricher Syntax-annotiertes Korpus
Z: `http://www.ifi.unizh.ch/cl/index5.html`, siclemat@ifi.unizh.ch (Simone Clematide)
F: Materialbasis für Maschinelle Sprachverarbeitung
S: Deutsch
M: geschrieben
G: 3000 Sätze
P: statisch
A: Morphosyntax, Syntax – wie NEGRA (manuell)
SB: Computerzeitung
V: auf Anfrage

4 Weiterführende Literatur

Wir wollen an dieser Stelle nicht auf weitere Literaturstellen verweisen, sondern auf Mailinglisten, bei denen Sie sich anmelden (*subskribieren*) können. Als angemeldeter Benutzer erhalten Sie alle an diese Listen gesendeten Beiträge. Durch diese Beiträge sind Sie über Entwicklungen in der Korpuslinguistik auf dem Laufenden. Sie können sich auch selbst beteiligen und dort Fragen stellen. Wenn Sie freundlich fragen, werden Sie in den meisten Fällen auch freundliche Antworten erhalten. Die Mailinglisten sind *Corpora* (`http://torvald.aksis.uib.no/corpora/`, folgen Sie dem Link zur *Info Page*, dort erfahren Sie, wie Sie Mitglied werden können), *Gesprächsforschung* (`http://www.gespraechsforschung.de/liste.htm`, mit online-Registrierung) und die *Linguist List* (`http://www.linguistlist.org/`).

aus: Lemnitzer/Zinsmeister (2006): S. 127-157

Wie man in den Wald hineinruft – Korpuslinguistik in der Praxis

Nach Lektüre dieses Kapitels werden Sie in der Lage sein, selbstständig korpusbasierte linguistische Untersuchungen entsprechend den genannten Vorbildern zu planen und durchzuführen.

1 Übersicht

In diesem Kapitel wollen wir linguistische Untersuchungen vorstellen, die auf deutschsprachigen Korpora basieren. Wir glauben, dass man aus diesen Beispiele etwas lernen kann – im positiven wie im negativen Sinn. Ein Blick in die germanistischen Fachzeitschriften zeigt, dass in den letzten fünf Jahren erstaunlich viele linguistische Arbeiten entstanden sind, denen Korpora zugrunde liegen. Diese Arbeiten sind freilich von recht unterschiedlicher Qualität, wie wir noch sehen werden. Sie sind auch thematisch weit gestreut.

Wir haben uns bei der Darstellung für die folgende Anordnung entschieden. Zunächst orientieren wir uns an den klassischen linguistischen Beschreibungsebenen: Orthographie, Morphologie und Wortbildung, Syntax, Lexikographie und Lexikologie. Bei einigen dieser Arbeiten, die sich nicht leicht in dieses Schema einordnen ließen, haben wir den Gegenstand als charakterisierendes Merkmal gewählt.

Der aufmerksame Leser wird sich über bestimmte Lücken in diesem Kapitel wundern. Gerade Disziplinen, in denen traditionell mit Korpora gearbeitet wurde, fehlen. Dies sind vor allem die Bereiche der historischen Linguistik und des Erstspracherwerbs. Auch das Gebiet der Computerlinguistik streifen wir nur kurz. Dies hat zwei Gründe. Zum einen fehlt uns auf den Gebieten der historischen Linguistik und der Psycholinguistik die nötige Kompetenz, um die schwierige und reichhaltige Materie kompakt darzustellen. Zum anderen wollen wir uns hier auf die innovativen Ansätze und Bereiche der linguistischen Analyse und Anwendung von Korpusdaten konzentrieren. Zu weiteren Ausführungen fehlt uns der Platz. Wir werden deshalb einige Aspekte auf der begleitenden Webseite ergänzen.

Eine kritische Würdigung der neueren korpuslinguistischen Arbeiten rundet dieses Kapitel ab.

2 Orthographie

Der deutschen Rechtschreibung liegt spätestens seit den Zeiten Konrad Dudens eine Norm zugrunde. Deshalb der Ausdruck *Recht*schreibung. Diese Norm, die allerdings auch Wandlungen und Reformen unterliegt, wird in der Schule vermittelt, sie kann nicht verhandelt werden. Sie ist allerdings öfter der Gegenstand von Diskussionen, was gerade die jüngste Rechtschreibreform gezeigt hat. Im fachlichen Diskurs der Linguistik stehen die Prinzipien hinter der Norm sowie Fragen ihrer Angemessenheit, Schlüssigkeit und Lernbarkeit zur Diskussion. Linguisten nehmen aktiv Anteil an der Weiterentwicklung dieser Norm, dabei werden sie von Zeit zu Zeit von interessierten Laien begleitet oder auch bekämpft.

Dementsprechend befasst sich das Gros der Arbeiten zur Rechtschreibnorm mit den folgenden Themen:

- Darstellung, Begründung oder Kritik der Norm;
- Präsentation der Norm, als Menge von Regeln und / oder als Liste von Einzelwörtern;
- Vermittlung der Norm im Sprachunterricht[1].

Diese Themen beziehen sich auf die festgesetzte Norm und nicht auf den tatsächlichen Sprachgebrauch. Empirische Untersuchungen an authentischen Sprachdaten sind hier überflüssig. Die Liberalisierung der orthographischen Norm im Zuge der letzten Rechtschreibreform macht es nun allerdings interessanter, am tatsächlichen Sprachgebrauch zu untersuchen, welche Varianten in welchen quantitativen Verhältnissen verwendet werden. Darüber hinaus gibt es bei einigen Textsorten orthographische Besonderheiten, die nicht Teil der Norm sind. Beide Aspekte der orthographischen Praxis sind Gegenstand jüngerer empirischer Untersuchungen.

Helmut Langner[2] untersucht den Wortschatz der Sachgruppe Internet auf morphologische, aber auch orthographische Besonderheiten. Er stellt fest, dass bei der Schreibung von Wörtern aus diesem Bereich orthographische Unsicherheiten deutlich werden: „Erstaunlich ist ...das starke Schwanken zwischen Zusammenschreibung und Schreibung mit Bindestrich, nicht selten sogar im selben Text ...Probleme

[1] Vgl. z.B. die Sammelbände von Augst (1997) sowie Eroms und Munske (1997).
[2] Vgl. Langner (2001). Seitenzahlenangaben in Klammern beziehen sich auf diesen Text.

haben Schreiber offensichtlich dann, wenn die Lexeme Konstituenten besitzen, die noch als fremdsprachig empfunden werden“ (105). Langner stützt seine Beobachtungen auf eine Belegsammlung, die er im Jahr 2000 aus verschiedenen Quellen, vor allem Zeitung und Rundfunk, zusammengestellt hat (97). Die Beobachtungen Langners zeigen, dass sich nicht alles in einer Rechtschreibnorm regeln lässt und manche Konzepte, wie das der Fremdworthaftigkeit mancher Ausdrücke, unscharf sind. Die reformierte Rechtschreibnorm trägt dem durch eine höhere Zahl an zugelassenen Varianten Rechnung. Dennoch wird es immer orthographische Probleme jenseits der Norm geben.

Christa Dürscheid untersucht zwei Typen von „Schreibungen, die in der Rechtschreibnormierung nicht geregelt sind“[3]. Es handelt sich dabei um die Binnengroßschreibung (z.B. *InterCity*) und um die Getrenntschreibung von Komposita (z.B. *Progamm Entwickler*). Ihre These lautet, dass sich in diesen Bereichen in der Sprachverwendung Tendenzen zeigen, die früher oder später die Rechtschreibnorm verändern werden. Sie stützt ihre Analysen auf unsystematisch gesammelte Belege aus verschiedenen Medien: Fernsehen, Radio, Zeitung, aber auch aus der Beschreibung von Software oder aus der Bahnwerbung. Am Schluss des Buches finden Sie hierzu eine Übung, die auf Material des *Wortwarte-Korpus* beruht.

In einer anderen Arbeit[4] untersucht Dürscheid Verstöße gegen die orthographische Norm an verschiedenen Textsorten, die Bestandteil computervermittelten Kommunikation sind. Die Daten, auf die sie diese Untersuchungen stützt, sind Mitschnitte von Chats sowie E-Mails. Ob die nicht-normgerechten Schreibweisen in der computervermittelten Kommunikation, die nicht auf technisches oder menschliches Versagen zurückzuführen sind, Auswirkungen auf die Schreibnorm und die Schreibpraxis außerhalb dieses Mediums haben werden, kann nicht vorausgesagt werden. Die Autorin fordert hierzu weitergehende empirische Untersuchungen. Dem kann man sich nur anschließen. Es wäre wünschenswert, wenn sich solche Untersuchungen auf ein öffentlich zugängliches Referenzkorpus der computervermittelten Kommunikation stützen könnten. Ein solches ist an der Universität Dortmund im Aufbau[5].

In verschiedenen Arbeiten, die um die Jahrtausendwende herum entstanden sind[6], werden vor allem graphostilistische Elemente in computervermittelter Kommunikation, und hier vor allem bei E-Mail und Chat, untersucht: Smileys, Sonderzeichen wie Stern (*) und at-Zeichen

[3] Vgl. Dürscheid (2000a), S. 223.
[4] Vgl. Dürscheid (2000b).
[5] Vgl. Kapitel 5.
[6] Vgl. Haase et al. (1997), Runkehl et al. (1998), Storrer (2000), Storrer (2001).

(@), prononcierte Großschreibung ganzer Wörter. Der Bereich ist für diese Formen der Kommunikation recht gut untersucht und auch solide korpuslinguistisch fundiert. Es wird in Zukunft zu zeigen sein, ob sich auch in Texten anderer neuer Medien, wie den über Mobiltelefone verbreiteten SMS, orthographische Sonderformen etablieren. SMS-Texte dürften allerdings wesentlich schwieriger zu akquirieren sein als Texte, die über das World Wide Web verbreitet werden[7].

3 Wortbildung

Die Wortbildung ist der kreativste Bereich einer tagtäglich verwendeten Sprache. Sprecher schaffen auf diese Weise unzählige neue Wörter, von denen viele nur dem einen, momentanen kommunikativen Zweck dienen und danach nie wieder verwendet werden.

Die Bausteine, aus denen im Deutschen neue Wörter geformt werden, sind:

- Wortstämme (z.B. *seh*, *Mutter*); eine Unterklasse der Stämme, die nicht selbständig ein Wort bilden können, wird *Konfix* genannt (z.B. *schwieger*, *thek*).
- Affixe, die nach ihrer Stellung zum Wortstamm unterschieden werden in Präfixe (z.B. *be-*), Suffixe (z.B. *-bar*) und Infixe (z.B. das Fugenmorphem *-s-*);
- Zwischen diesen beiden Klassen stehen Elemente, die sich von selbständigen Wortstämmen zu Affixen entwickeln, unter Verlust eines eigenen semantischen Gehalts (z.B. *-mäßig* in Wörtern wie *gefühlsmäßig*). Diese Bausteine werden in der neueren Literatur *Affixoide* genannt.
- Flexive, die grammatische Merkmale eines Worts wie Kasus oder Tempus markieren (z.B. *-en*, das als Flexiv die Infinitivform und die erste und dritte Person Plural eines Verbs markieren kann).

Ziel der Wortbildungsforschung als linguistischer Disziplin ist es, die Regeln und Beschränkungen zu formulieren, denen die freie Kombination dieser Bausteine unterliegt, und die Merkmale der aus der Kombination der Bausteine entstehenden Wortbildungsprodukte zu beschreiben. Zum Beispiel

- darf das Suffix *-bar* nur mit verbalen Wortstämmen kombiniert werden. Das entstehende Wort wird als Adjektiv verwendet. Der Beitrag des Suffixes zur Gesamtbedeutung des Adjektivs ist es meist, dass die durch den verbalen Stamm beschriebene Handlung dem Gegenstand auf den das neue Adjektiv sich bezieht, als Potenzial zugeschrieben wird (X ist ableitbar → X kann abgeleitet werden).
- muss in manchen Fällen zwischen die zwei Bestandteile eines Kompositums ein Fugenmorphem treten. Die Notwendigkeit des Fugenelements wird phonologisch begründet, es macht den Übergang vom letzten Phonem des ersten Wortstamms zum ersten Phonem des zweiten Wortstamms leichter (z.B. *Arbeit-s-amt*, *Tag-e-bau*).

Die Wortbildung als produktiver Prozess des Sprachausbaus steht im Spannungsverhältnis zum Lexikon einer Sprache. Wenn täglich Hunderte von neuen Wörtern gebildet werden, dann kann das Lexikon einer Sprache oder eines einzelnen Sprechers niemals vollständig in Hinblick auf das Vokabular der Sprache sein. Es ist deshalb ähnlich wie in der Syntax eine wichtige linguistische Aufgabe, die Regeln zu beschreiben denen dieser kreativ Prozess unterliegt[8]. Diese Regeln steuern die Produktion neuer Wörter und ermöglichen es den Hörern, diese neuen Wörter korrekt zu interpretieren[9].

Empirische Sprachdaten sind auch für Wortbildungsforschung wichtig:

- Große Korpora enthalten viele Belege für die meisten Wortbildungsmuster und durchweg mehr Beispiele, als ein Wörterbuch verzeichnen kann. Gerade die nicht in Wörterbüchern verzeichneten, kontextuell gesteuerten Gelegenheitsbildungen bilden einen wichtigen Prüfstein für theoretische Annahmen zu Regeln, Regularitäten und Beschränkungen in der Wortbildung;
- Viele Wortbildungsprodukte werden erst verständlich und interpretierbar, wenn man den Kotext sieht, in dem das Wort verwendet wird. Besonders Komposita bedürfen oft der Stützung durch den Kotext[10].

[7] Es gibt dennoch einige korpusbasierte Arbeiten zu diesem Thema, z.B. Schwitalla (2002), Doering (2002), allerdings beziehen sich diese Arbeiten nicht auf die Themen Rechtschreibnorm und Rechtschreibpraxis.

[8] Dass in diesem Teil der Sprache Regeln wirken, sieht man an Bildungen wie *un kaputtbar*, die deshalb so auffällig sind, weil sie gegen diese Regeln verstoßen. In unserem Beispiel ist das Ziel des Regelverstoßes, Aufmerksamkeit zu erregen, und dies ist sicher gelungen.

[9] Oftmals ist dafür aber auch ein größerer Kotext oder Kontext erforderlich, wie das Beispiel *BVB-Transfer* zeigt. Ob ein BVB transferiert wird oder ein BVB etwas transferiert, erschließt sich, wenn man weiß oder erfährt, dass der BVB ein Fußballverein ist, der seine Mannschaft durch Transfers von Spielern verändert.

[10] Auf den Zusammenhang hat kürzlich Corinna Peschel in ihrer Monographie zum Verhältnis von Wortbildung und Textkonstitution hingewiesen, vgl. Peschel (2002)

3.1 Qualitative Aspekte

In den letzten Jahren ist eine Reihe von korpusbasierten Fallstudien zu einzelnen Wortbausteinen erschienen. Hierzu gehören Arbeiten von Angelika Feine sowie von Anke Lüdeling und Stefan Evert zur nichtmedizinischen Verwendung von *-itis*-Kombinationen[11], eine Arbeit von Nikolaus Ruge zum Suffixoid *-technisch*[12], eine Studie zur Valenz der *be*-präfigierten Verben von Piklu Gupta[13] sowie ein Aufsatz von Annette Klosa zu Verben mit dem Präfix *gegen-*[14]. Mehr korpusgestützte Arbeiten zu Details der Wortbildung erscheinen uns wünschenswert.

Wir wollen in diesem Abschnitt exemplarisch die Arbeit von Susanne Riehemann zur Beschreibung der Adjektive mit dem Suffix *-bar* vorstellen[15]. Riehemann versucht anhand von intensiven Korpusrecherchen die Wortbildungsregeln und -beschränkungen im Zusammenhang mit der Verwendung des Suffixes *-bar* zu erfassen und in der Lexikonkomponente des Grammatikformalismus *Head-Driven Phrase Structure Grammar* (HPSG) zu beschreiben (2-3). Ihre Arbeit ist damit sowohl für die theoretische Linguistik als auch für die Computerlinguistik von Interesse.

Riehemann stützt ihre Untersuchungen auf neun Korpora, ein großes und acht kleinere, mit insgesamt knapp 18 Millionen laufenden Wörtern (Token). Die Frequenzangaben zu den *-bar*-Adjektiven bezieht die Autorin ausdrücklich nur auf das mit 10,7 Millionen Token größte Korpus, das Zeitungskorpus des Instituts für deutsche Sprache in Mannheim. Die kleineren Korpora bezeichnet sie als zu wenig repräsentativ, um quantitative Aussagen darauf zu stützen (5). Im einzelnen untersucht sie die folgenden Aspekte:

- Die Klassen von *-bar*-Ableitungen, vor allem hinsichtlich der zugrunde liegenden Verben. Riehemann berücksichtigt die Frequenzverteilung dieser Adjektive, die das typische Profil aller produktiven sprachlichen Prozesse aufweist: es gibt wenige hochfrequente Wörter, die weit über die Hälfte aller vorkommenden Wörter ausmachen, und sehr viele selten vorkommende Wörter (9-12);

[11] *Handyritis, Aufschieberitis* etc., vgl. Feine (2003) und Lüdeling und Evert (2004). Auf die Arbeiten von Lüdeling und Evert werden wir im nächsten Abschnitt genauer eingehen.

[12] Vgl. Ruge (2004), interessant sind hier weniger die transparenten Bildungen wie *verfahrenstechnisch*, sondern vielmehr neudeutsche Bildungen wie *gefühlstechnisch*.

[13] Vgl. Gupta (2000).

[14] Vgl. Klosa (2003), die Untersuchungen basieren auf dem Korpora des Instituts für deutsche Sprache und auf dem DWDS-Korpus.

[15] Vgl. Riehemann (1993). Die Seitenzahlen in Klammern verweisen auf diesen Text.

- die Form und Funktion der Wortbildungsprodukte, also der so entstandenen Adjektive, wobei sie vor allem deren syntaktische (mögliche Komplemente der Adjektive) und semantische Eigenschaften betrachtet (5-9);
- in einem weiteren Abschnitt diskutiert Riehemann syntaktische, semantische und pragmatische Beschränkungen des Wortbildungsprozesses, die erklären, warum einige Bildungen ungrammatisch sind, wohingegen andere, ebenfalls vom prototypischen Muster – mit einem transitiven Verb als Basis – abweichende Wörter durchaus bildbar sind (z.B. *abbaubar* mit einem intransitiven Verb als Basis und *verformbar* mit einem reflexiven Verb als Basis.) (12-16);
- Riehemann zieht auch die Argumente der zugrunde liegenden Verben in Betracht, die von dem abgeleiteten Adjektiv ‚ererbt' werden (*Ein Auto nach Deutschland importieren* → *Ein nach Deutschland importierbares Auto*). Vor allem bei der Bestimmung von Beschränkungen hinsichtlich der Vererbung von Argumenten erweist sich der Blick in das Korpus als sehr hilfreich (17-19);
- schließlich beschreibt Riehemann Unterschiede im attributiven und prädikativen Gebrauch dieser Adjektive.

Im zweiten, dem Hauptteil der Arbeit entwickelt Riehemann eine formale Beschreibung der lexikalischen Eigenschaften dieser Adjektivgruppe im Rahmen eines HPSG-Lexikons, die all den im ersten Teil der Arbeit beschriebenen Generalisierungen gerecht wird. Die Arbeit endet mit zwei Anhängen, in denen zum einen alle im Korpus vorgefundenen *-bar*-Adjektive, zum anderen die häufigsten 300 Adjektive in der Reihenfolge ihrer Häufigkeit aufgelistet sind (70-78). Riehemanns Arbeit ist ein wichtiger Beitrag zu einer formalen Beschreibung von Wortbildungsprozessen am Beispiel des vermutlich produktivsten Suffixes der deutschen Sprache. Weitere Arbeiten in diesem Stil sind wünschenswert.

3.2 Qualitativ-quantitative Aspekte

In jüngster Zeit ist in verstärktem Maße die Produktivität von Wortbildungselementen, wie z.B. dem Suffix *-bar*, untersucht worden. Die Produktivität in der Wortbildung hat einen qualitativen und einen quantitativen Aspekt. Beide erfordern unterschiedliche Analysemethoden.

- Der qualitative Aspekt hängt zusammen mit der Menge der Elemente, mit denen ein bestimmtes Morphem kombiniert werden kann. So ist z.B. der Anwendungsbereich des Suffixes *-bar* auf verbale Basen beschränkt, und hier fast ausschließlich auf die transitiven Verben. Das Suffix *-sam* hingegen tritt zusammen mit verbalen Basen

(*arbeit-sam*) und mit adjektivischen Basen (*selt-sam*) auf. Der Anwendungsbereich von *-bar* und damit die Menge der hiermit bildbaren Wörter ist also beschränkter als der Anwendungsbereich von *-sam*;

- der quantitative Aspekt der Wortbildung kann informell beschrieben werden als die Wahrscheinlichkeit, mit der man einem mit einem bestimmten Morphem gebildeten neuen Wort begegnet, nachdem man bereits eine bestimmte Anzahl von Wörtern beobachtet hat. In einer anderen Sichtweise wird der Produktivitätsindex bestimmt von der relativen Anzahl der Wörter, die bisher nur einmal in den beobachteten Daten auftauchten[16]. In dieser Interpretation wird man nach Analyse eines Korpus der deutschen Gegenwartssprache feststellen, dass das Suffix *-bar* relativ produktiv ist, die Produktivität des Suffixes *-sam* hingegen gegen null tendiert. Mit anderen Worten, die Wörter mit dem Suffix *-sam* sind vollständig aufzählbar.

Wie man an den obigen Beispielen sieht, sind der qualitative und der quantitative Aspekt der Produktivität von Wortbildungselementen unabhängig voneinander. Die qualitative Analyse kann anhand einer Belegsammlung durchgeführt werden. Für die quantitative Analyse ist die Analyse eines kompletten, möglichst großen Korpus allerdings zwingend notwendig. Dies hat zwei Gründe:

- Erstens kann man im Hinblick auf Vorkommenshäufigkeiten von Wörtern oder Wortbildungsmustern weder die eigene Intuition noch die Intuition anderer Muttersprachler zu Rate ziehen. Hinsichtlich quantitativer Verhältnisse ist unser Sprachgefühl zu unzuverlässig;
- zweitens muss man für die hier zur Diskussion stehende Analyse eine große Menge von Texten sukzessive nach der Anzahl und Häufigkeit der Vorkommen eines bestimmten Musters durchforsten.

Anke Lüdeling und Stefan Evert[17] untersuchen den quantitativen Aspekt der Produktivität des Suffixes *-lich*. Sie verwenden hierfür ein Zeitungskorpus von ca. 3 Millionen laufenden Wörtern. Die Analyse der Klasse aller mit *-lich* gebildeten Wörter ergibt ein ziemlich unscharfes Bild. Die Analyse wird aber präziser, nachdem die Autoren vier verschiedene Klassen gebildet haben: a) *-lich* mit adjektivischer Basis (z.B. *grün-lich*), b) *-lich* mit verbaler Basis (z.B. *vergess-lich*), c) *-lich* mit nominaler Basis (z.B. *ärzt-lich*) und d) *-lich* mit phrasaler Basis (z.B. *vorweihnacht-lich*). Die Kombination des Suffixes mit nominaler Basis

[16] Eine formale Beschreibung dieses als *Vocabulary Growth Curve* bezeichneten Phänomens gibt Baayen (2001).
[17] Vgl. Lüdeling und Evert (2003).

ist sehr produktiv, die Kombination mit verbaler Basis hingegen unproduktiv. Für die beiden anderen Bildungsmuster ist die Datenmenge zu gering für eine ausreichend genaue Bewertung. Die Autoren zeigen weiterhin, dass es auch unter den Nomen herausragend produktive Stämme gibt (z.B. *X-geschicht-lich*), die eine weitere Klassifizierung der Nomen nahe legen. Wie man an diesem Beispiel sieht, kann die qualitative Analyse von der quantitativen Analyse profitieren. Letztere fungiert sozusagen als Lackmustest für die Güte einer qualitativ begründeten Klassifizierung.

Anke Lüdeling, Stefan Evert und Ulrich Heid[18] zeigen aber auch, dass der automatischen Analyse von Korpora im Hinblick auf Anzahl und Häufigkeit von Wortbildungsmustern Grenzen gesetzt sind. Dies hängt mit der Fehleranfälligkeit der Analysemöglichkeiten zusammen, die eine manuelle Durchsicht der Daten beim heutigen Stand der Technik erforderlich machen. Probleme bereiten:

- Tippfehler in den Texten;
- Wörter, die zufällig mit der gleichen Zeichenkette wie das Suffix enden (z.B. *Balsam, Sesam*);
- Wörter, die scheinbar eine Derivation sind, im Grunde aber eine Komposition mit einem früher derivierten Wort (z.B. *Kadaverge horsam* → *Kadaver+Gehorsam*, nicht jedoch → *Kadavegehor-sam*) Beide Fälle sind mit den heutigen Mitteln morphologischer Analyse nicht zu unterscheiden. So wurde z.B. *unverzichtbar* gebildet durch Präfigierung von *verzichtbar*; *befahrbar* wurde gebildet durch Suffigierung von *befahren*. Nur das letzte Wort ist relevant für die die Wortbildung mit *-bar*[19].

Lüdeling und Evert zeigen das Potenzial, aber auch die Grenzen einer korpusgestützten Produktivitätsanalyse beim heutigen Stand der Technik[20]. Die Relevanz solcher Untersuchungen liegt in den folgenden Anwendungsgebieten:

- In der Lexikographie kann man sich bei unproduktiven Wortbildungselementen auf die Auflistung der wichtigsten lexikalischen Einheiten beschränken. Für produktive Wortbildungselemente ist der Ansatz eines eigenen Artikels zu erwägen, in dem die Verwendungsregularitäten erklärt werden sollten;
- im Sprachunterricht spielt die Vermittlung der morphologischen und semantischen Regularitäten produktiver Wortbildungselemente ein

[18] Vgl. Lüdeling et al. (2000) und Evert und Lüdeling (2001).
[19] Die Beispiele entstammen Evert und Lüdeling (2001).
[20] Die Notwendigkeit manueller Intervention ist einer der Gründe, warum die Autoren für ihre *-lich*-Studie ein relativ kleines Korpus gewählt haben.

wichtige Rolle. Es ist wahrscheinlich, dass Lerner Wörtern dieses Bildungstyps begegnen werden, die nicht im Wörterbuch stehen[21].

4 Syntax

In der Syntaxforschung kann man drei Arten der Verwendung von Korpora unterscheiden: Erstens die Suche nach einzelnen Beispielen bzw. Gegenbeispielen im Rahmen einer bestimmten Theorie. Zweitens die Erhebung von Frequenzangaben zu bestimmten Phänomenen, oft im Rahmen eines Vergleichs von konkurrierenden syntaktischen Alternativen. Der dritte Typ ist eine syntaxbasierte Auswertung von Korpora, dessen Ergebnisse zwar Konsequenzen für eine zu Grunde gelegte syntaktische Theorie haben, aber auch als Datenbasis für weitere Untersuchungen in anderen Bereichen der Linguistik oder Computerlinguistik genutzt werden können.

Quer zu dieser Einteilung liegt die Frage nach der Art der Korpusabfrage, welche stark mit der zur Verfügung stehenden Annotation der Korpora gekoppelt ist. Wird nur über Wortformen gesucht oder auch über Wortarten oder wird sogar eine weiterführende Annotation genutzt? In jedem Fall ist die Verwendung eines Suchwerkzeugs, das mit regulären Ausdrücken arbeitet, extrem hilfreich[22]. Detmar Meurers und Stefan Müller[23] diskutieren eine Reihe von Fallbeispielen, in denen sie Korpusanfragen zu syntaktischen Phänomenen durchspielen. Sie erläutern anschaulich, wie man die linguistische Fragestellung in Konzepte der Korpusannotation übersetzen kann. Siehe hierzu auch Abschnitt 3 im vierten Kapitel.

Mangels verfügbarer Ressourcen haben Syntaktiker bisher oftmals nur mit wortbasierter Suche recherchiert z.B. Pittner (1999) oder Ehrich (2001) auf den IDS-Korpora. Das wird sich in Zukunft wahrscheinlich ändern, nachdem inzwischen vollständig syntaktisch annotierte Baumbanken wie *TüBa-D/Z* und *TIGER* zur Verfügung stehen. Letztere umfasst in ihrem zweiten Release vom Dezember 2005 immerhin 900 000 Token und hat damit eine beachtliche Größe erreicht.

[21] Korpusbasierte morphologische Analysen spielen auch in der Computerlinguistik und hier besonders in der Computerlexikographie eine Rolle. Korpusanalysen dienen hier dazu, das Regelhafte und das Idiosynkratische zu trennen: alles, was nicht in Regeln gefasst werden kann, muss in Lexika beschrieben werden. Eine wichtige Rolle spielen hier die Arbeiten im Umfeld des morphologischen Lexikons *IMSLex*, vgl. Fitschen (2004). Wir können auf diesen Aspekt an dieser Stelle nicht näher eingehen und verweisen auf die computerlinguistische Fachliteratur.

[22] Siehe den Exkurs zu den *Regulären Ausdrücken* in Kapitel 4 auf S. 90.

[23] Vgl. Meurers (2005), Meurers und Müller (in Vorb.).

Ein zweites Problem ist die Handhabung der Korpora: Sie zu lizei zieren und auf dem eigenen Computer zu installieren, ist nicht jede manns Sache. Es besteht aber die Tendenz, dass in Zukunft immer mel Ressourcen auch online zugänglich gemacht werden, so dass sich au die technischen Hürden verringern werden.

Im Folgenden stellen wir Ihnen zu den drei eingangs genannten T pen stellvertretend ein paar Arbeiten vor. Ein Beispiel für den erst Typ, die Suche nach Beispielen, sind die Arbeiten von Stefan Müll zur mehrfachen Vorfeldbesetzung. Er verwendet für seine Recherch die IDS-Korpora über die Online-Anfrage COSMAS, das Material, d auf den *DigiBib*-CDs[24] zur Verfügung steht, und die Tageszeitung *t* (persönliche Auskunft). Das Ergebnis seiner Recherche sind Beispie wie[25]:

(1) [Öl] [ins Feuer] goß gestern das Rote-Khmer-Radio:...

Hier stehen zwei unabhängige Konstituenten vor dem finiten Verb i Vorfeld (*Öl* bzw. *ins Feuer*). Die Belegsammlung zeigt die Natürlic keit des Phänomens. Müller argumentiert, die Häufigkeit des Auftrete zeige, dass man die Daten, deren Existenz in der theoretischen Liter tur teilweise bestritten wurde, nicht einfach ignorieren kann. Müll selbst schlägt eine Analyse im Rahmen der *Head Driven Phrase Str ture Grammar* (HPSG) vor[26]. Die empirische Untersuchung macht d Vielfalt des Phänomens deutlich, wobei sich gewisse Muster in den D ten feststellen lassen[27]. Das Phänomen weist aber auch Beschränkung auf, vgl. die Ungrammatikalität von Beispiel (2). Die empirischen Dat helfen, Kontexteigenschaften zu identifizieren, die die weitere Analy unterstützen.

(2) * Maria Max gab ein Buch.

In einer methodisch ähnlichen Arbeit untersucht Gabriele Kniffka Syntax und Pragmatik von NP-Aufsplittung im Deutschen (im Ra men der so genannten DP-Hypothese der generativen Grammatik) Die Belege geschriebener Sprache stammen bei ihr aus verschieden

[24] Vgl. DigiBib: `http://eris.hbz-nrw.de/`.

[25] Quelle: taz, 18.06.1997.

[26] Konkret nimmt er an, dass die Konstituenten im Vorfeld durch ein abstraktes V lizenziert sind, vgl. Müller (2005).

[27] Siehe Müller (2003) und Müllers Belegsammlung auf `http://www.cl.uni-breme de/~stefan`.

[28] Vgl. Kniffka (1996).

Druckerzeugnissen, zusätzlich wertet sie aber auch ein kleines Korpus der gesprochenen Sprache aus[29].

Angelika Storrer[30] untersucht die Distribution von Nominalverbgefügen (NVG) wie *Unterricht erteilen*. Ein relativ allgemeines Verb (*erteilen*) tritt zusammen mit einer Nominalisierung als Objekt (*Unterricht*) in fester Wendung auf[31]. Storrer vergleicht die Verteilung der NVGs mit denen des jeweiligen Basisverbs (hier *unterrichten*). Motivation für diese Arbeit ist die immer wieder zu lesende Behauptung, dass die NVG nur eine phrasale Umschreibung des Basisverbes sei – und zudem ein schlechter Sprachstil. Anders als die bisher genannten Arbeiten wertet Storrer ein spezifisches Korpus aus, das DWDS-Kernkorpus. Sie analysiert die Belege zunächst qualitativ und untersucht dabei vergleichend das semantische und kombinatorische Potenzial von NVG und Basisverb, z.B. mögliche Selektionsrestriktionen oder Modifikationsmöglichkeiten am Basisverb und an der Nominalisierung. Letztere bietet eine Reihe von Optionen, die beim Basisverb nicht gegeben sind, wie die Modifikation durch bestimmte Adjektive, durch Relativsatz oder Spezifikator sowie bestimmte Koordinationsmöglichkeiten. Belege wie (3) im Kontrast mit dem konstruierten (4) können als Gegenbeispiel zur ‚Umschreibungsthese' gewertet werden.

(3) ... dem Krieg eine Absage erteilen.

(4) ??dem Krieg absagen.

Eine zusätzliche quantitative Auswertung zur wechselseitigen Paraphrasierbarkeit ergibt, dass die Basisverben mehrdeutig (*polysem*), die entsprechenden NVGs hingegen spezifischer sind und meist nur eine der Bedeutungen des Basisverbs tragen. Die NVG erlaubt es demnach Ambiguitäten zu vermeiden. Zum Beispiel ist *unterrichten* ambig zwischen den Lesarten *mitteilen* und *lehren*, während *Unterricht erteilen* nur die eine Bedeutung hat. Das Fazit der Studie ist, dass Nominalverbgefüge keine ‚semantischen Dubletten' des Basisverbs sind – die oben erwähnte Stilfrage stellt sich damit nicht. Storrers Arbeit leitet direkt zum zweiten Verwendungstyp über, dem der Frequenzerhebung.

[29] Jan-Philipp Soehn sammelt ebenfalls Belege aus diversen Quellen und entwickelt darüber eine HPSG-Analyse zu idiomatischen Wendungen, vgl. (Soehn, 2006). Wir erwähnen die Arbeit, weil sie eine interessante Datensammlung für weitere Untersuchungen auf CD bereithält.

[30] Vgl. Storrer (2006a).

[31] Die Klasse der Nominalverbgefüge ist in sich nicht homogen. Storrer (2006b) differenziert hier weiter und stellt einen korpusbasierten Vergleich von zwei Unterklassen vor.

Die beiden folgenden Arbeiten erheben Frequenzdaten auf ein syntaktisch annotierten Korpus. Sie sind beide an der Distribution Relativsätzen interessiert und verbinden die Untersuchung der Korp frequenz mit psycholinguistischen Experimenten. In der ersten Arl untersuchen Uszkoreit et al.[32], welche Faktoren einen Einfluss dai haben, ob ein Relativsatz adjazent, d.h. direkt benachbart, zu seir Bezugsnomen steht oder extraponiert im Nachfeld auftritt.

(5) Er hat [das Buch, [das er gestern erst gekauft hat],] heute gele

(6) Er hat [das Buch] heute gelesen, [das er gestern erst gekauft h

Sie verwenden eine Vorstufe des NEGRA-Korpus mit 12 000 vollstär syntaktisch annotierten Sätzen, welches sich aber als zu klein erw so dass sie auf ein weiteres Korpus zurückgreifen. Die Untersuch kann damit auf einer Textbasis von 1 Millionen Wörtern durchgefü werden[33]. Das Ergebnis der quantitativen Studie legt eine performa orientierte Erklärung der Distribution nahe. Bestimmend sind die F toren *Distanz* (zwischen Bezugsnomen und potenzieller extraponie Position) und *Länge* (Gewicht des Relativsatzes in Wortanzahl). ne ähnliche Auswertung, diesmal auf dem kompletten NEGRA-Kor wird von Schade et al.[34] durchgeführt. Sie suchen nach geschachtel Relativsätzen in der geschriebenen Sprache und finden Beispiele (Klammerung wurde hinzugefügt):

(7) Er hat jene Heiterkeit, [die ein Tierlehrer, [der an sich auf Pfe dressuren geeicht ist], braucht], um auch ein so spaßiges Spel kel wie den „Schweizer Bergbauernhof" durchzustehen.

Um einen Eindruck von der spontanen Produktion zu bekommen, v ten sie auch die Verbmobil-Baumbank zur gesprochenen Sprache au Dort finden sie keine geschachtelten Relativsätze, sondern nebenge nete Strukturen wie (Klammerung wiederum hinzugefügt):

(8) Ja, also erstmal zum Hotel: Da haben wir noch drei verschied Hotels, [die wir Ihnen anbieten können], [die noch Zimmer haben].

[32] Vgl. Uszkoreit et al. (1998).

[33] Das zweite Korpus ist nur POS-annotiert und erfordert, wie die Autoren bemer viel zeitaufwändige Handarbeit in der Auswertung.

[34] Vgl. Schade et al. (2003).

[35] Die Verbmobil-Baumbank ist 2005 als *TüBa-D/S* veröffentlicht worden.

Die beiden Korpusstudien verwenden Schade et al. als Ausgangsbasis für ihre weiterführenden psycholinguistischen Experimente zur Relativsatzperzeption.

Eine Arbeit, die in den dritten Verwendungsbereich fällt, also die syntax-basierte Korpusauswertung für weitere Anwendungen, stellt Nadine Aldinger[36] vor. Sie entwickelt auf der Basis einer halbautomatischen Textanalyse Regeln, die man einsetzen kann, um verschiedene Lesarten von Genitivattributen deverbaler Nomen zu unterscheiden. Das klingt kompliziert, gemeint sind damit Beispiele wie:

(9) (...) die Bodenmessungen des städtischen Umweltamtes (...)

(10) (...) Vermietung ganzer Etagen an polnische Landarbeiter (...)

Das Interessante hierbei ist, dass der jeweilige postnominale Genitiv (*des städtischen Umweltamtes* bzw. *ganzer Etagen*) beim zugrunde liegenden Verb (hier also *messen* und *vermieten*) unterschiedliche Funktionen einnehmen kann (*das Umweltamt* als Subjekt und *ganze Etagen* als Akkusativobjekt). Zu erkennen, um welche Lesart es sich handelt, ist eine wichtige Grundlage für verschiedene Anwendungen in der Computerlinguistik, z.B. für die Informationsextraktion. Aldinger verwendet das *Frankfurter Rundschau Korpus* am Institut für Maschinelle Sprachverarbeitung in Stuttgart. Das Korpus umfasst 40 Millionen Wörter, ist lemmatisiert, automatisch getaggt und (rekursiv) gechunkt. Sie extrahiert mit Hilfe der Abfragesprache *Corpus Query Processor* (CQP) Beispiele, die durch die vorhandene Annotation automatisch nach folgenden Merkmalen sortiert werden können: Beim deverbalen Kopfnomen (hier *Bodenmessung* bzw. *Vermietung*) speichert Aldinger z.B. Numerus, Definitheit, Kasus, ggf. den Spezifikator (Wort und Wortart) und die adjektivischen Modifikatoren sowie den Nicht-Kopfanteil bei Komposita (*Boden* von *Bodenmessungen*); bei einer dem Genitiv nachgestellten PP die Präposition und den Kasus der eingebetteten NP. Für die Genitiv-NP selbst notiert sie u.a. Numerus, Definitheit und das Kopflemma. Wir geben die Liste der Kontextfaktoren so detailliert wieder, um klar zu machen, dass die für die Interpretation des Genitivs verantwortlichen komplexen Zusammenhänge der Faktoren nur empirisch festzustellen sind, d.h. nur durch eine quantitative Studie aufgedeckt

[36] Vgl. Aldinger (2005).

werden können[37]. Diese Art von Daten entziehen sich der Introspektion[38].

Timm Lichte[39] arbeitet ebenfalls mit einem (rekursiv) gechunkten Korpus, der TüPP-D/Z. Er verwendet 2.7 Millionen Sätze des Gesamtkorpus, um automatisch Negative Polaritätselemente (NPI)[40] zu identifizieren. NPIs sind Ausdrücke, die nur im Umfeld von bestimmten negativen Ausdrücken und Fragekontexten lizenziert sind wie *(nicht) ganz geheuer*. Lichte legt die Annahme zu Grunde, dass sich NPIs und ihre Lizenzierer wie Kollokationen verhalten. Außer der Menge der Lizenzierer gelten alle anderen Lemmata des Korpus als potentielle NPIs[41] Sein System erstellt eine Rangliste der Lemmata, die manuell überprüft werden muss. Unter den obersten 20 Kandidaten findet man schöne Beispiele wie *verdenken*, *unversucht*, *umhin* oder *lumpen*. Lichte zeigt auch auf, wie seine Methode auf Mehrwort-NPIs erweitert werden kann. In einem Experiment dazu erhält er Kandidaten wie *unversucht lassen*, *ganz geheuer*, *umhin zu kommen* oder *lumpen lassen*.

5 Computerlinguistik

Die Computerlinguistik ist ein Bereich, in dem Korpora eine wichtige Rolle spielen. Zunächst dienen sie einfach als Datenquelle für das empirische Arbeiten. Der Computerlinguist sichtet Korpusdaten, um seine Hypothesen, Modelle oder Programme an authentischem Material zu entwickeln und zu prüfen. In diesem Vorgehen unterscheidet er sich nicht von anderen Linguisten.

Der Unterschied besteht darin, dass der Computerlinguist die Korpora auch in großem Maßstab zum Entwickeln und Prüfen seiner Programme nutzen kann. Was ist damit gemeint?

Bei der Entwicklung von Programmen nutzt er die Frequenzinformationen, die in einem Korpus stecken, z.B. beim *Training* von statistischen Programmen[42]. Diese Programme beinhalten Regeln, deren Anwendungen über so genannte Gewichte gesteuert werden. Eine Regel mit höherem Gewicht wird bevorzugt angewendet. Die Werte für die

[37] Siehe z.B. die *Multivariate Analysis* in McEnery und Wilson (2001, S. 88f.).

[38] Eine methodisch sehr ähnliche Arbeit wird von Kathrin Beck in (Beck, 2006) durchgeführt. Sie wertet Kontextfaktoren für die Interpretation von Präpositionalergänzungen von *ung*-Nominalisierungen in der TüBa-D/Z aus.

[39] Vgl. Lichte (2005).

[40] *Negative Polarity Item* (NPI).

[41] Lichte beschränkt die Untersuchung auf Lemmata, die häufiger als 40 mal im Korpus vorkommen. Er erhält damit eine Ausgangsmenge von fast 35 000 Lemmata.

[42] Im vierten Kapitel hatten wir Ihnen im Exkurs zum Part-of-Speech Tagging z.B. das Training des Brill-Taggers vorgestellt.

Gewichte werden aus Korpora abgeleitet, indem man die Wahrscheinlichkeiten für die Regeln anhand eines Korpus ermittelt (in der Computerlinguistik sagt man, das Programm *lernt* die Wahrscheinlichkeiten beim *Training*). Stark vereinfacht zählt das Programm dabei, wie oft eine Regel bei der Analyse des Korpus angewendet wird[43].

Ein Beispiel für das Lernen aus Korpora ist die *Grammatikinduktion*. Aus den Annotationsstrukturen des Korpus werden Frequenzen für Grammatikregeln abgelesen. Im Extremfall leitet man sogar die Grammatikregeln selbst aus dem Korpus ab (Anette Frank[44] erzeugt z.B. eine lexikalisierte *Tree Adjoining Grammar* auf der Basis des NEGRA-Korpus).

Das Training kann auch nur indirekt auf einem Korpus stattfinden. Manchmal werden zuerst Daten aus einem Korpus extrahiert und zum Beispiel in einer Datenbank gesammelt. Die im Abschnitt zur Syntax beschriebenen Arbeiten von Nadine Aldinger und Timm Lichte sind Beispiele dafür. Aldinger sammelt komplexe syntaktische und morphologische Informationen zu Genitivergänzungen von *-ung*-Nominalisierungen, um die Lesarten der Ergänzungen vorherzusagen. Lichte listet Kookkurrenzen von Wörtern und Lizenzierern für Negative Polaritätselemente auf, um mit Hilfe eines statistischen Programms Kandidaten für Negative Polaritätselemente zu bestimmen.

Sabine Schulte im Walde[45] zeigt, wie man mit computerlinguistischen Methoden die Verbklassen von Levin (1993) auf deutschen Daten nachvollziehen kann. Sie trainiert zunächst eine Grammatik auf dem Huge German Corpus, um Frequenzinformationen über Verben, deren Argumentrahmen und die aufgetretenen nominalen Realisierungen der Argumente zu erfassen. In einem zweiten Schritt entwickelt sie ein Programm, das aus diesen Informationen Klassen von Verben bilden kann (das Programm *clustert* die Verben in Gruppen), z.B.[46]:

(11) Verben, die sich auf eine Basis beziehen:
basieren, beruhen, resultieren, stammen

(12) Verben der Maßänderung:
reduzieren, senken, steigern, verbessern, vergrößern, verkleinern, verringern, verschärfen, verstärken, verändern (...)

[43] Zwei empfehlenswerte englischsprachige Einführungen zur statistischen Sprachverarbeitung sind Jurafsky und Martin (2000) und Manning und Schütze (1999).
[44] Vgl. Frank (2001).
[45] Vgl. Schulte im Walde (2003).
[46] Wir stellen hier nur korrekte Beispiele vor, um das Ergebnis zu veranschaulichen. Das Programm clustert teilweise auch Verben in eine Gruppe, die keine gemeinsame Bedeutung besitzen.

Eine weitere Verwendungsweise von Korpora in der Computerlinguistik ist das Testen von Programmen, anders ausgedrückt die *Evaluierung*. Hierzu benötigt man ein linguistisch annotiertes Korpus (den *Gold Standard*), das idealerweise mit den Strukturen annotiert ist, die das Programm erzeugen soll. Der Idealfall ist allerdings nicht immer gegeben, da – wie Sie ja inzwischen wissen – Annotation sehr aufwändig und kostenintensiv ist. Man muss manchmal Kompromisse eingehen und z.B. die Ausgabe des eigenen Programms auf das vorgegebene Format des Testkorpus abbilden. Letzteres hat den einen Vorteil, dass man auf diese Art verschiedene Programme unmittelbar anhand desselben Testkorpus vergleichen kann. Wenn man testet, muss man sich klar machen, dass auch das Testkorpus Fehler enthalten kann. Es bietet sich daher an, als obere Grenze bei einer Evaluierung nicht 100% Übereinstimmung zu verlangen, sondern sich an der Übereinstimmung der Annotatoren des Gold Standards zu orientieren (am *Inter Annotator Agreement*).

6 Lexikologie und Lexikographie

Der Nutzen von Korpora für die Lexikographie ist vielfältig, was an anderer Stelle ausführlich beschrieben wird[47]. Wir wollen uns hier auf eine Zusammenfassung aus der Sicht des lexikographischen Prozesses und auf einige Felder beschränken, die auch für das Deutsche gut bearbeitet wurden.

Aus der Sicht des lexikographischen Prozesses[48] werden Korpora in den folgenden Phasen konsultiert:

- Bei der Wörterbuchplanung, besonders bei der Finanzplanung, spielen die Existenz und die Verfügbarkeit von Korpora für den durch das Wörterbuch zu beschreibenden Gegenstand eine Rolle. Wichtig sind auch die Werkzeuge, die die für die Lexikographen relevanten Informationen aus den Korpora extrahieren und präsentieren. Hier ist möglicherweise Entwicklungs- und Anpassungsarbeit notwendig.
- Korpora können wichtige Hinweise für die Lemmaauswahl geben. So kann die Häufigkeit, mit der eine lexikalische Einheit in einem Korpus vorkommt, darüber entscheiden, ob sie in die Lemmaliste eines Wörterbuchs aufgenommen wird oder nicht[49].

[47] Vgl. Engelberg und Lemnitzer (2001), Wiegand (1998) und die dort erwähnte Literatur sowie, für das Englische, Ooi (1998).
[48] Vgl. hierzu vor allem Kapitel 6 in Engelberg und Lemnitzer (2001).
[49] Ausführlich hierzu Scholze-Stubenrecht (2002).

- Den Hauptteil lexikographischer Arbeit bildet das Erstellen der Wörterbuchartikel zu den Lemmata. Bei einem allgemeinsprachlichen Standardwörterbuch müssen die lexikalischen Zeichen auf allen linguistischen Ebenen beschrieben werden. Hierfür bilden Korpora eine Informationsquelle. Betrachten wir ein Beispiel. Es muss beschrieben werden, ob bestimmte Verben, die mentale Zustände ausdrücken – *wissen, glauben, meinen* etc.
  - mit *dass*-Sätzen und *ob*-Sätzen als Ergänzung verwendet werden können; wenn dies der Fall ist
  - welches, wenn beide Ergänzungen möglich sind, die häufigere Variante ist oder ob eine der beiden Varianten sehr selten ist, und weiter
  - ob die Verwendung der Ergänzungen auf bestimmte Kontexte beschränkt ist, z.B. negative Kontexte oder bestimmte Zeitformen des Verbs:

  (13) *Ich weiß, ob das geht.

  (14) Ich weiß *nicht*, ob das geht.

  (15) *Er wusste, ob das geht.

  (16) Er *wird* schon wissen, ob das geht[50].

  Diese subtilen Unterscheidungen können am besten durch die gründliche Analyse eines Textkorpus ermittelt werden.
- Korpora stellen eine wichtige Quelle von Verwendungsbeispielen dar. Lexikographen können auf Grund ihrer Sprachkompetenz zwar Beispiele erfinden, es hat sich aber erwiesen, dass diese bei weitem nicht an die Qualität von Korpusbelegen heranreichen[51].
- Die Häufigkeit ihrer Verwendung kann ein wichtiges Kriterium für die Anordnung von Lesarten in einem Artikel für ein sprachliches Zeichen sein. Vor allem in Lernerwörterbüchern sollte das Häufige vor dem Seltenen erscheinen oder das Seltene sogar unerwähnt bleiben, je nach Umfang des Wörterbuchs.
- Ein wichtiger Aspekt der Verwendung lexikalischer Zeichen ist ihre Verwendung in typischen Kotexten. Manche lexikalischen Zeichen tauchen in nur einem oder sehr wenigen Kotexten auf (z.B. *Hehl, fackeln*), viele lexikalische Zeichen treten typischerweise mit ein kleinen Anzahl anderer lexikalischer Zeichen auf und bilden mit di sen Kollokationen oder idiomatische Wendungen (typische Begleit von *hart* sind z.B.: *Bandagen, Droge, Leben, Währung*). Statistisc Verfahren, auf großen Korpora angewendet, geben Auskunft üb diese typischen Paarungen. Auch hier sind Korpora der sprachlich Intuition selbst der erfahrensten Lexikographen überlegen.
- In den Produktionsphasen nach der Erstellung der Wörterbuchi tikel – Korrektur und Drucklegung – spielen Korpora naturgem eine geringe Rolle. Einzelne Entscheidungen in der Korrekturph se können bei Bedarf an Korpora überprüft werden. In der Phi der Materialsammlung zwischen zwei Auflagen eines Wörterbuc kommt Texten, die nach der Drucklegung der letzten Auflage schienen sind, wieder eine größere Bedeutung zu.

[50] Wir empfehlen Ihnen, in einem Wörterbuch ihrer Wahl nachzuschlagen und zu prüfen, ob Sie auf die Fragen, die wir hier gestellt haben, eine Antwort finden. Wenn Sie Muttersprachler sind, versetzen Sie sich in die Situation eines Nichtmuttersprachlers, der diese Verben korrekt verwenden möchte. Oder machen Sie den Test mit einem Wörterbuch einer anderen Sprache.

[51] Luise Pusch hat eine lesenswerte Satire geschrieben, für die sie reichlich Beispiele der von den Duden-Redakteuren produzierten Belegprosa verwendet, vgl. Pusch (1984).

Die Werkzeuge, die Lexikographen typischerweise für diese Arbeit v wenden, sind Programme für die quantitative Analyse von Korpo um z.B. die Verwendungshäufigkeit bestimmter lexikalischer Zeiche insgesamt oder in bestimmten Lesarten – oder typische Kombinat nen sprachlicher Zeichen zu ermitteln. Des Weiteren werden Progra me verwendet, die für ein bestimmtes lexikalisches Zeichen alle V kommenskontexte in einer vom Lexikographen festlegbaren Anordnu präsentieren[52]. Die Kombination dieser Werkzeuge hilft, aus dem M der Texte durch Auswahl und Filterung der Daten den Lexikograph die Informationen zu liefern, die sie für ihr Handwerk der lexikalisch Beschreibung benötigen[53].

Wir werden uns im Folgenden auf drei Felder konzentrieren, auf nen die germanistische Korpuslinguistik bereits einige Erfolge erziel d.h. interessante und relevante Ergebnisse zu Tage fördern konnte. D sind die Lexikonbereiche der Neologismen und Anglizismen sowie Kombination einzelner lexikalischer Zeichen in Kollokationen und f ten Wendungen.

6.1 Kollokationen und Phraseme

Als Kollokation wird das gemeinsame Vorkommen zweier sprachlicl Zeichen miteinander bezeichnet. Ein Element einer Kollokation tritt

[52] Diese Werkzeuge präsentieren ‚Keywords in Context', und werden deshalb KW Tools genannt, die Daten, die sie erzeugen, *Konkordanzen*. Wir gehen in Absch 3 näher darauf ein.

[53] Ein Desiderat sind allerdings immer noch Werkzeuge, die automatisch die Bel auswählen, in denen ein Schlüsselwort in einer bestimmten Lesart verwendet w Dies ist ein Forschungsgegenstand der Computerlinguistik.

Umfeld des anderen Teils auf. So kommt im vorletzten Satz z.B. *als* im Umfeld von *Kollokation* vor, *sprachlicher* im Umfeld von *Zeichen* etc. Wichtig ist, dass dieses gemeinsame Vorkommen nicht zufällig ist. Nun kann man mit Recht behaupten, dass die Wahl eines Wortes in einem durchdachten Text niemals zufällig ist. Wir müssen es also etwas anders formulieren. Wir sprechen von einer Kollokation, wenn ein lexikalisches Zeichen ein anderes lexikalisches Zeichen als Kotext bestimmt, meist unter Ausschluss anderer, bedeutungsähnlicher Zeichen. Der Charakter dieser Auswahl wird deutlich, wenn wir einige in etwa gleichbedeutende Wortverbindungen in verschiedenen Sprachen betrachten. In Tabelle 10 haben wir einige Paare zusammengestellt.

Sprache 1	Sprache2	Wörtliche Übersetzung
Schlange stehen	sp: hacer cola	Schlange machen
sich die Zähne putzen	fr: se laver les dents	sich die Zähne waschen
den Tisch decken	en: lay the table	den Tisch legen
dichtes Haar	en: thick hair	dickes Haar
harte Währung	fr: devises fortes	starke Währung

Tabelle 10: Kollokationen in verschiedenen Sprachen

Man sieht an den Daten in Tabelle 10, dass

- die Auswahl eines Wortes durch ein anderes arbiträr und zugleich in einer Einzelsprache konventionalisiert ist, es sich also bei Kollokationen um komplexe sprachliche Zeichen handelt;
- die Auswahl eines Wortes durch ein anderes sich nicht regelhaft beschreiben lässt. Man *putzt* sich die Zähne und *wäscht* sich die Haare oder Hände, man ist mit etwas *hoch* zufrieden oder über etwas *stark* enttäuscht oder gar von etwas *voll* genervt. Diese Wortverbindungen müssen deshalb als Ganzes gelernt bzw. im Wörterbuch gesucht werden.

Als Kollokation im weiteren Sinn hat man im Umfeld des Kontextualismus jedes gemeinsame Vorkommen zweier Wörter im gleichen Kotext bezeichnet[54]. Dieser sehr weite Begriff wird bereits im Umfeld des Kontextualismus weiter eingegrenzt, zunächst auf die Wortpaare, die üblicherweise miteinander vorkommen[55]. Sidney Greenbaum berücksichtigt zudem die syntaktischen Relationen zwischen den miteinander vorkommenden Wörtern[56]. So könnten die Beziehungen zwischen den miteinander vorkommenden Wörtern der Wortklassen Nomen und Adjektiv oder Nomen und Verb gezielt untersucht werden. Die Verbindung von *Als* und *Kollokation* aus unserem obigen Beispiel würde sich dagegen nicht als Kollokation qualifizieren.

Franz Josef Hausmann schließlich führt den Unterschied zwischen Basis und Kollokator ein. Zwischen diesen beiden Elementen besteht eine gerichtete Beziehung; die Basis bestimmt den Kollokator. Welche Konsequenzen für die Lexikographie das hat, wollen wir an dem Beispiel der Kollokation *schütteres Haar* erläutern. Wenn ein Sprecher oder Schreiber einen Text produzieren möchte, dann ist ihm daran gelegen zu erfahren, welche Prädikate dem Gegenstand *Haar* sprachlich zugeschrieben werden können (z.B. *lang, kurz, blond, rot, braun, graumeliert, strähnig, voll, dicht, schütter*). Dieser potenzielle Benutzer eines Wörterbuchs wird bei der Basis (*Haar*) nachschlagen, um Unsicherheiten bei der Wortwahl zu klären. Hausmann geht es in erster Linie um die Verbesserung der lexikographischen Praxis, die in Einklang zu bringen sei mit den unterschiedlichen Nachschlagebedürfnissen von Benutzern, die einen Text verstehen, und Nutzern, die einen Text erstellen wollen[57]. Wir teilen Hausmanns Meinung, dass es sinnvoll ist, dem Begriff *Kollokation* ein schärferes Profil zu geben. Für sprachtechnologische Zwecke aber mag es genügen, die Wortpaare zu finden, die häufiger als erwartbar miteinander vorkommen. Um beiden Phänomenen gerecht zu werden, wollen wir hier zwischen *Kookkurrenz* und *Kollokation* (im engeren Sinn) unterscheiden.

- Als *Kookkurrenz* soll das gemeinsame Vorkommen zweier Wörter in einem gemeinsamen Kotext betrachtet werden. Die Länge des betrachteten Kotextes kann als Textfenster einer bestimmten Länge festgelegt werden. Im Allgemeinen wird vom einzelnen Beleg abstrahiert und das gemeinsame Vorkommen zweier Wörter in vielen Kotexten betrachtet werden. Es kann zudem die Reihenfolge des Auftretens beider Wörter in den Belegen als unterscheidendes Kri

[54] „[...] innerhalb der britischen Schule des Kontextualismus [...] wurde unter *Kollokation* das faktische Miteinandervorkommen zweier oder mehrerer beliebiger Wörter und/oder lexikalischer Einheiten [...] verstanden [...]. Der Terminus *Kollokation* war in der Theorie des Kontextualismus an keinerlei normative Bewertung hinsichtlich Korrektheit oder Grammatikalität der untersuchten Wortverbindungen gekoppelt.", vgl. Lehr (1996), S. 2.

[55] „By collocation is meant the *habitual* association of a word in a language with other particular words in sentences." (Robins 1964, zit. nach Lehr (1996), S. 5).

[56] „A more valuable, if more modest, contribution might be made to the study of collocations if a relatively homogenous class of items were selected and an investigation undertaken of the collocation of each item in the class with other items that are related syntactically in a given way.", vgl. Greenbaum (1970), S. 13.

[57] Zu dieser Position vgl. vor allem Hausmann (1985) und Hausmann (2004).

terium zweier Kookkurrenzen festgelegt werden[58]. Ferner kann festgelegt werden, dass die Wörter einer Kookkurrenz häufiger (im gegebenen Textfenster) miteinander vorkommen, als dies der Fall wäre, wenn die Wörter zufällig verteilt wären. Man spricht in diesem Fall von einem *signifikanten* Kovorkommen beider Wörter und verwendet statistische Assoziationsmaße, um dies zu messen[59].

- Eine *Kollokation* muss natürlich den oben genannten Kriterien genügen, darüber hinaus aber auch eine innere Struktur, in Form einer Hierarchie zwischen Kollokationsbasis und Kollokator aufweisen. Darüber hinaus müssen die Glieder einer Kollokation in einer syntaktischen Beziehung zueinander stehen, z.B. als Köpfe einer Verbalphrase und einer gleich- oder untergeordneten Nominalphrase, oder als Kopf einer Nominalphrase und Kopf einer untergeordneten Adjektivphrase[60].

Es ist offensichtlich, dass Korpora für das Aufspüren von Kookkurrenzen und Kollokationen von großem Nutzen, wenn nicht gar unverzichtbar sind. Je größer das Korpus, desto mehr Belege für ein beliebiges Wortpaar wird man darin finden. Dies macht die darauf basierenden Statistiken zuverlässiger. Im einfachsten Fall, dem der Kookkurrenz, reicht es, das Korpus in eine Menge von Textfenstern aufzuteilen und zu ermitteln: a) in wie vielen Festern Wort_1 und Wort_2 gemeinsam vorkommen, b) in wie vielen Fenstern nur Wort_1 vorkommt, c) in wie vielen Fenstern nur Wort_2 vorkommt und d) in wie vielen Fenstern weder Wort_1 noch Wort_2 vorkommen. Die meisten Assoziationsmaße setzen diese vier Werte bzw. ihre Summen miteinander in Beziehung. Das Ergebnis der Anwendung eines Assoziationsmaßes auf ein Wortpaar ist eine Kennziffer, durch die dieses Wortpaar mit anderen Wortpaaren in Beziehung gesetzt werden kann. Wortpaare mit hohen Kennziffern sind signifikante Kookkurrenzen und damit gute Kandidaten für Kollokationen. Die anderen Bedingungen für eine Kollokation müssen allerdings auch gegeben sein. Um dies zu prüfen, braucht man ein Korpus, bei dem zumindest die Wortarten annotiert sind, oder eine Belegsammlung.

Elisabeth Breidt wendet ein solches Verfahren auf ein wortartenannotiertes Korpus an, um Nomen-Verb-Kollokationen zu ermitteln[61]. Lothar Lemnitzer[62] experimentiert mit verschiedenen Assoziationsmaßen

[58] Die Wortfolge *doch eben* bedeutet eben doch etwas anderes als die Wortfolge *eben doch*.

[59] Eine Übersicht über statistische Assoziationsmaße geben Lemnitzer (1997), Kapitel 4, und Evert (2004).

[60] Einige Beispiele für diese Beziehungen befinden sich in Tabelle 10.

[61] Vgl. Breidt (1993).

[62] Vgl. Lemnitzer (1997), Kap. 4.

und arbeitet ebenfalls mit einem wortartengetaggten Korpus, beläs es aber bei einer Fallstudie, den Kollokanten des lexikalischen Zeiche *hart*.[63]. Joachim Wermter und Udo Hahn extrahieren Kollokation zwischen Präpositionalphrasen und Verben aus einem großen, ebe falls wortartengetaggten Korpus[64]. Von hoher praktischer Relevanz si schließlich auch die Arbeiten am Institut für maschinelle Sprachver beitung der Universität Stuttgart. Stellvertretend sei hier auf die Arb von Heike Zinsmeister und Ulrich Heid hingewiesen[65]. Die Autoren e trahieren aus einem getaggten und partiell geparsten Zeitungskorp Kombinationen von Verb, Nomen und modifizierendem Adjektiv, tre nen die relevanten von den irrelevanten Kombinationen und klass zieren die relevanten Tripel halbautomatisch in sechs Klassen, die d Spektrum von der idiomatischen Wendung (z.B. *offene Türen einre nen*) bis zur gänzlich freien Fügung (z.B. *konkrete Zahlen nennen*) e decken. Die Relevanz dieser Arbeit für die praktische Lexikographie offensichtlich. Die Autoren diskutieren auch die Grenzen und Proble ihres Ansatzes. So gibt es zur Zeit kein Verfahren, das auf der Ba der Unterschiede der sechs Klassen eine vollständige und vollkomme Klassifizierung erreichen kann[66].

Es ist davon auszugehen, dass die Extraktion von Kollokationen u anderen mehrwortigen Lexemen zur Alltagspraxis in den großen de schen Wörterbuchredaktionen gehört, es ist aber nicht zu erwarten, d man von dort Interessantes über diese Arbeit erfahren wird. Es ble zu hoffen, dass die Korpusarbeit sich positiv auf die Qualität der W terbücher gerade in diesem Bereich auswirkt[67].

Die korpusbasierte Untersuchung von festen Redewendungen, *Ph seme* genannt, steht hinter der Untersuchung von Kollokationen bisl deutlich zurück. Eine Ausnahme bildet eine Arbeitsgruppe um Chr tiane Fellbaum an der Berlin-Brandenburgischen Akademie der W senschaften. Diese Gruppe hat es sich zum Ziel gesetzt, systematis möglichst vollständig und mit synchroner und diachroner Perspekt

[63] Vgl. Lemnitzer (1997), Kap. 5. Das Hauptziel dieser Arbeit ist es, korpusgestü Mehrwortlexeme zu ermitteln, Kollokationen sind dort nur ein Aspekt unter mel ren.

[64] Vgl. Wermter und Hahn (2004). Zwei ihrer Beispiele sind *unter Druck geraten* u *in den Griff bekommen*.

[65] Vgl. Zinsmeister und Heid (2003).

[66] Einige Beispiele für alle sechs Klassen werden in Zinsmeister und Heid (200 Tabelle 5, präsentiert.

[67] Hausmann kritisiert u.E. völlig zu Recht die Einordnung vieler Kollokationen ter dem Stichwort des Kollokanten und zudem das Fehlen vieler Kollokationen Duden Stilwörterbuch, einem Wörterbuch also, das vor allem auf die Benutzung Produktionswörterbuch angelegt ist, vgl. Hausmann (2004), S. 310.

die Gruppe der aus einer Verbalphrase und einer untergeordneten Nominalphrase bestehenden Phraseme zu untersuchen[68]. Phraseme zeichnen sich dadurch aus, dass

- sie nach der Grammatik der entsprechenden Sprache nicht immer wohlgeformt sind (z.B. *ganz Ohr sein*);
- sie semantisch intransparent sind, die einzelnen Bestandteile also nicht die Bedeutung haben, die sie in freier Verwendung haben (z.B. *die Katze aus dem Sack lassen*);
- sie nur begrenzt modifizierbar sind (vgl. *einen Kater haben*, *einen furchtbaren Kater haben*, *einen grau gescheckten Kater haben*, im letzten Fall geht die idiomatische Lesart – unter den Folgen überhöhten Alkoholgenusses leiden – verloren)[69].

Die von Fellbaum und ihrem Team untersuchten verbalen Phraseme zeichnen sich dadurch aus, dass sie oft komplexe Sachverhalte benennen und deshalb nicht einfach in die semantischen Strukturen des Lexikons einer Sprache eingefügt werden können[70].

In einer detaillierten Arbeit untersuchen sie die Funktion der hochgradig unspezifischen Pronomen *etwas* und *ein(en)* als Ergänzungen verbaler Phraseme[71]. In manchen Fällen haben diese Ergänzungen Argumentstatus und referieren auf etwas, wenn auch sehr Unspezifisches (z.B. *etwas auf der hohen Kante haben*). In anderen Fällen hat *etwas* keinen Argumentstatus und referiert nicht (z.B. *jemandem etwas husten*). Die Autoren vermuten, dass der „Platzhalter" hier grammatische Funktionen übernimmt. Zum einen ermöglicht er die Einführung eines indirekten Objekts (das die Existenz eines direkten Objekts voraussetzt; *etwas* füllt diesen Platz aus). Zum anderen erzwingt *etwas* die Interpretation des Verbs und damit des gesamten Phrasems als zeitlich eingegrenztes Ereignis. Zwischen diesen beiden Verwendungen von *etwas* gibt es, wie die Autoren zeigen, etliche Zwischenstufen. Ähnliche Befunde werden bei der Analyse von *ein(en)* ermittelt.

Die Arbeit ist vor allem für die lexikographische Praxis relevant. Da die beiden Hauptfunktionen von *etwas* und *ein(en)* die möglichen Modifikationen des Phrasems im Gebrauch beeinflussen, sollten bei der lexi-

[68] Vgl. Fellbaum (2002), Abschnitt 6.

[69] Für eine detaillierte Analyse vgl. Keil (1997).

[70] *einen zwitschern* ist eben mehr als eine bestimmte Art zu trinken, das Phrasem evoziert eine ganze Szene, bei der das Trinken alkoholischer Getränke eine Rolle spielt. Dieses „mehr" ist es, was die Forscher vor allem interessiert, vgl. Fellbaum (2002), Abschnitt 3.

[71] Vgl. Fellbaum et al. (2004).

kographischen Ansatzform diese beiden Elemente zumindest graphis unterschieden werden[72].

6.2 Neologismen

Im weitesten Sinne sind Neologismen sprachliche Zeichen, also Wört Bedeutungen und Wendungen, die zu einem bestimmten Zeitpunkt v den Sprechern, die sie verwenden, als neu empfunden werden.

Neologismen können von ihrer Form her unterteilt werden in Neu xeme und Neubedeutungen. Das Wort *Podcast* ist vor nicht allzu lan Zeit als ein Neulexem in den deutschen Sprachgebrauch aufgenomm worden, da es diese Wortform im Deutschen Lexikon bisher nicht gab Das Wort *Maus* hingegen erhielt in den frühen siebziger Jahren ei Neubedeutung, es bezeichnet seitdem ein Peripheriegerät am Com ter.

Neologismen können weiterhin an Hand des Grades ihrer Lexil lisierung und ihrer Integration in den deutschen Sprachgebrauch u terschieden werden. Danach bezeichnen Neologismen im engeren Si Wörter, die weitgehend lexikalisiert sind. Sie werden relativ häufig u bereits über einen längeren Zeitraum verwendet und in die Neuaul gen allgemeinsprachlicher Wörterbücher aufgenommen. Hierzu geh sicher das Verb *simsen* (=eine SMS verschicken). Daneben gibt es Gelegenheitsbildungen, die nur ein oder wenige Male verwendet werd danach wieder in Vergessenheit geraten und auch nicht in Wörterbüc aufgenommen werden. Ein Beispiel hierfür ist das Wort *semimerk* (womit eine Frisur im Stil von Angela Merkel bezeichnet wurde). Di sogenannten *Okkasionalismen* sind von der Lexikographie und Lexi logie lange Zeit als uninteressant abgetan worden. Sie bieten aber die Wortbildungsforschung und für die Lexikographie interessantes M terial[74]. Entlang dieser letzten Unterscheidung haben sich zwei Forn der Neologismenlexikographie herausgebildet:

- Die *aktuelle Neologismenlexikographie* sammelt und archiviert W ter vom ersten Augenblick ihres Erscheinens an. Diese Sammlun enthalten zwangsläufig viele Okkasionalismen, da zum Zeitpunkt ersten Erscheinens eines Wortes nicht vorhergesagt werden kann, dieses Wort sich im Gebrauch etablieren wird. Erfahrene Lexikog phen können lediglich gute Voraussagen über die Entwicklung ei

[72] Vgl. Fellbaum et al. (2004), Abschnitt 5.

[73] *Podcast* bezeichnet die meist private Distribution von Hörbeiträgen, im Stile e Radiosenders, über das World Wide Web.

[74] Vgl. hierzu Peschel (2002).

Wortes treffen. Ein Beispiel für die aktuelle Neologismenlexikographie ist die *Wortwarte*[75].

- Die *retrospektive Neologismenlexikographie* sammelt und beschreibt in Spezialwörterbüchern dieses Lemmatyps die Wörter, die im Beschreibungszeitraum aufgekommen sind und sich bereits etabliert haben. Ein Beispiel hierfür ist das Wörterbuch *Neuer Wortschatz. Neologismen der 90er Jahre im Deutschen*[76]. Dementsprechend wird hier der Begriff *Neologismus* im engeren Sinn verwendet.

Korpora spielen in der Neologismenforschung und -lexikographie die folgenden Rollen:

- Bei regelmäßiger Beobachtung zum Beispiel der Tagespresse lässt sich mit einiger Sicherheit feststellen, wann ein Wort (in einer bestimmten Bedeutung) zum ersten Mal verwendet wird.
- Die quantitative Auswertung eines größeren Korpus, das den Sprachgebrauch eines bestimmten Zeitraums repräsentiert, ergibt, welche Wörter ausreichend oft belegt sind, so dass man von einem etablierten Wort, also einem Neologismus im engeren Sinn sprechen kann.
- Anhand eines zeitlich gegliederten Korpus lässt sich auch ermitteln, welche Wortbildungselemente eine wachsende Rolle bei der Bildung neuer Wörter spielen. So ist z.B. das Präfix *Cyber-* erst seit Ende des letzten Jahrzehnts in Verwendung und gehört seitdem zu den produktiven Wortbildungselementen.
- In Korpora belegte Verwendungsgewohnheiten geben Auskunft über sich verfestigende Eigenschaften des Gebrauchs, z.B. die Zuordnung eines Genus zu einem aus dem englischen entlehnten Wort.
- Schließlich liefern Korpora Belege, die als Vorlagen für den Erwerb des normgerechten Gebrauchs eines neuen Wortes wichtig sind.

Linguistische und lexikographische Neologismus-Forschung ist also ohne die Analyse authentischer Sprachdaten unmöglich. Für lange Zeit war die manuelle Analyse und Auswertung von Printwerken die einzig machbare Arbeitsmethode, und vor allem in der Wörterbucherstellung werden neue Wörter noch heute überwiegend auf diese Art gesammelt. Es gibt aber Projekte, in denen digitalisierte Korpora für diese Zwecke genutzt werden. Ein Beispiel hierfür ist die *Wortwarte* von Lothar Lemnitzer und Tylman Ule. Seit Ende 2000 werten die Autoren regelmäßig die Online-Ausgaben mehrerer Tages- und Wochenzeitungen aus. Die Wörter dieser Texte werden mit der Wortliste eines Referenzkorpus abgeglichen. Die nach diesem Abgleich übrig gebliebenen Wörter werden

[75] Im WWW unter der Adresse `www.wortwarte.de` erreichbar.
[76] Vgl. Herberg et al. (2004).

täglich manuell ausgewertet. Pro Tag werden im Durchschnitt 15 neue Wörter ausgewählt, beschrieben und mit einem Beleg aus der Fundstelle versehen. Alle neuen Wörter werden regelmäßig mit einer Frequenzangabe versehen, die auf der Zahl der von der Suchmaschine *Google* gelieferten Treffer zu dem jeweiligen Suchwort basiert. Neben dem online zugänglichen Wörterbuch mit mittlerweile über 20 000 Einträgen stehen alle Wortlisten zur Verfügung. Mit diesen Daten lassen sich z.B. Aussagen über Tendenzen der Wortbildung treffen[77]. Auch in diesem Projekt wird mit einem weiten Begriff von *Neologismus* gearbeitet, der auch Gelegenheitsbildungen umfasst. Zweitens wird in diesem Projekt, und dies ist ein neuer Ansatz, versucht, das Web als kontinuierliche Quelle aktueller Sprachdaten zu nutzen.

Das einzige größere Spezialwörterbuch des Lemmatyps Neologismen, das der retrospektiven Neologismenlexikographie verpflichtet ist, bildet die vom Institut für deutsche Sprache herausgegebene Sammlung *Neuer Wortschatz. Neologismen der 90er Jahre* von Dieter Herberg, Michael Kinne und Doris Steffens[78]. Bei der Erstellung dieses Wörterbuchs wurde mit einem engeren Neologismusbegriff gearbeitet. Gegenstand des Wörterbuchs sind die Neuwörter und Neubedeutungen, die „in den 90er Jahren des 20. Jahrhunderts in der deutschen Allgemeinsprache aufgekommen sind, sich darin ausgebreitet haben, als sprachliche Norm allgemein akzeptiert und in diesem Jahrzehnt von der Mehrheit der deutschen Sprachbenutzer über eine gewisse Zeit hin als neu empfunden wurden."[79] Das Erscheinungsjahr des Wörterbuchs, 2004, deutet an, dass die Autoren zwar zeitlich relativ nah an ihrem Beschreibungsgegenstand sind, aber doch weit genug entfernt, um den Prozess der Lexikalisierung aus der Rückschau beobachten zu können. Als Primärquelle des Werks diente ein Teil der IDS-Korpora, das Texte des untersuchten Zeitraums umfasst. Dazu kam eine Wortkartei mit ca. 10 000 Einträgen[80].

Etwa 700 Neologismen wurden aus diesen Quellen ausgewählt und bearbeitet, wobei die Korpusbefunde „die Grundlage für die Darstellung zahlreicher Datentypen in den Wortartikeln"[81] bilden, z.B. der Anga-

[77] Die Einträge sind auf der Website der Wortwarte, `www.wortwarte.de`, veröffentlicht, die Seite wird täglich aktualisiert. Auf der Website befinden sich auch weitere Informationen zum Projekt. Die Wortlisten können bei den Autoren angefordert werden.
[78] Vgl. Herberg et al. (2004).
[79] S. XXIII. Besonders das letzte Kriterium steht auf empirisch schwachen Füßen, es ist zu vermuten, dass das Sprachgefühl der Autoren hier repräsentativ für das Sprachgefühl aller Sprachbenutzer gesetzt wird.
[80] S. XVI f.
[81] S. XVI.

ben zu Flexion, zu Wortbildungsmustern und zu den Verwendungsweisen. Inwiefern sich ein solches lemmabezogenes Spezialwörterbuch neben aktuellen allgemeinsprachlichen Wörterbüchern, vor allem dem Rechtschreibduden, etablieren wird, bleibt abzuwarten. Das Buch ist jedenfalls eine interessante Quelle für die Sprachlehre bei fortgeschrittenen Lernern. Vielleicht ergeben sich aus dieser konsequent korpusbezogenen Arbeit auch Impulse für die traditionelle Lexikographie des Deutschen und deren Produkte.

Schließlich sollen noch einige Spezialarbeiten zu Neologismen aus linguistischer Sicht, und hier vor allem die Beiträge von Hilke Elsen zu Neologismen in einigen Varitäten des Deutschen, erwähnt werden[82].

Mit den beschriebenen Projekten hat sich eine linguistische und lexikographische Praxis der Analyse von Neologismen auch des Deutschen etabliert. Neu sind vor allem die Nutzung des World Wide Web als Datenquelle und die stärkere Berücksichtigung von Okkasionalismen.

6.3 Anglizismen

Anglizismen sind ein weiterer markierter Bereich des deutschen Wortschatzes. Unter dem Begriff Anglizismus versteht man alle aus dem Sprachkontakt einer Sprache mit dem Englischen resultierenden Phänomene der Entlehnung und der Beeinflussung des Sprachsystems der Zielsprache[83]. Aus vielerlei Gründen ist das Englische nach 1945 zur stärksten Gebersprache im linguistischen Kontakt geworden. Lexikalische Einheiten aus dem britischen und vor allem dem amerikanischen Englisch bilden einen nicht zu vernachlässigenden Teil des Vokabulars der deutschen Sprache, der mehr oder weniger stark in die deutsche Sprache integriert ist. Das Englische ist auch eine wichtige Quelle für Neologismen. Anglizismen stellen das System und vor allem den Gebrauch der deutschen Sprache vor besondere Schwierigkeiten.

- Orthographisch weicht die Norm der Getrennt- und Zusammenschreibung sowie der Bindestrichschreibung von der englischen Norm und orthographischen Praxis ab[84].
- Die Aussprache kann sich eher am englischen Original orientieren (z.B. *Banker* [bæŋkə] anstatt [baŋkə] oder *kiten* [kaɪtn̩] anstatt [ki:tn̩]) oder am phonologischen System des Deutschen (z.B. bei *Download* wird die zweite Silbe eher als [lo:t] gesprochen mit deutscher Auslautverhärtung anstatt des ursprünglichen [ləʊd]).

[82] Vgl. Elsen (2002), Elsen (2004) und Elsen und Dzikowicz (2005).
[83] Vgl. Bartsch (2002), S. 312.
[84] Vgl. hierzu, aus dem Blickwinkel der alten Rechtschreibnorm, Augst (1992).

- Morphologisch ergeben sich Probleme bei der Pluralbildung (*Flye* -> *?Flyers* oder *?Flyer*) und der Konjugation (*?geuploaded, ?upge loaded*).
- Die größten Probleme entstehen beim Genus, das im Englische nicht festgelegt ist (der / die / das *Engine, Toolbar, Airbag*?).
- Grammatisch ergeben sich die geringsten Probleme, da die System sich hier sehr ähneln (heißt es *Aktien traden* oder *mit Aktien traden* letzteres in Analogie zu *handeln*).
- Weiterhin bringen Anglizismen Unsicherheiten in der Verwendun mit sich – *Searchengine* wird man wahrscheinlich nicht im Gespräc mit der Großmutter verwenden und *abchillen* nicht im Gespräch mi dem Chef.

Wie man sieht, müssen die Verwendungsbedingungen von entlehnte Wörtern erst im Prozess der Entlehnung ausgehandelt werden, beson ders dort, wo sie in der Gebersprache nicht ausreichend spezifizier sind[85]. Die Integration in das sprachliche System des Deutschen kan unterschiedlich weit fortschreiten (vgl. *Majonäse* oder *Kode*, im Gegen satz dazu ist der Ausdruck *Computer* kaum integriert). Sie wird vo Normen wie etwa der zur Rechtschreibung gesteuert, und die Aufnahm eines Anglizismus in die Wörterbücher des Deutschen geht mit Fest legungen der Verwendungsnorm auf den verschiedenen linguistische Ebenen einher.

Anglizismen werden bevorzugt in drei Wörterbuchtypen aufgenom men:

- Spezialwörterbücher des Lemmatyps Anglizismus. Hier ist vor allen das sprachdokumentarische *Wörterbuch der Anglizismen* von Car stensen und Busse zu nennen[86]. Es gibt aber auch einige sprachpu ristisch ausgerichtete Werke auf diesem Regalbrett, z.B. das *Wör terbuch überflüssiger Anglizismen* von Bartzsch[87].
- Fremdwörterbücher, in denen die aus anderen Sprachen entlehnte oder aus dem Griechischen und Lateinischen überkommenen lexika lischen Einheiten versammelt sind, deren Gebrauch in der Alltags sprache weniger üblich ist (z.B. *Parallaxe, Chintz*).
- Allgemeinsprachliche Standardwörterbücher wie das Duden Univer salwörterbuch oder Spezialwörterbücher z.B. zur Rechtschreibung.

[85] Die nicht vorhandene Genusmarkierung bei englischen Nomen ist hierfür ein Bei spiel.
[86] Vgl. Carstensen und Busse (1993). Die lexikographische Arbeit stützt sich auf da Paderborner Korpus, im Wesentlichen eine Belegsammlung, sowie die Korpora, di Mitte der achtziger Jahre am Institut für deutsche Sprache zur Verfügung standen vgl. Carstensen und Busse (1993), S. 47-53.
[87] Vgl. Bartzsch (2004).

Normunsicherheit besteht vor allem bei Wörtern, die noch nicht in Wörterbüchern registriert sind. Im Prinzip sollten hier die generellen orthographischen und grammatischen Normen des Deutschen hinreichend präzise Richtlinien für den Gebrauch geben. Augst zeigt jedoch, dass zumindest die Regeln der (alten) Rechtschreibung nicht ausreichen und selbst in den Wörterbüchern bei einzelnen lexikalischen Einheiten inkonsequent angewendet wurden[88]. Auch die Regeln der reformierten Rechtschreibung erleichtern es nicht, die korrekte Schreibung eines Anglizismus zu erschließen, wie Jürgen Dittmann und Christian Zitzke zeigen[89]. Die Autoren zeigen weiterhin in einer korpusbasierten Studie, dass in einigen Bereichen der Sprachgebrauch deutlich von den Normen, der offiziellen wie auch der der Nachrichtenagenturen, abweicht[90]:

- Bei rein englischen Komposita dominiert die Getrenntschreibung, eine deutliche Abweichung von beiden Normen (z.B. *Key Accounter*, *Call Center*);
- Bei den Mischkomposita mit englischen und deutschen Bestandteilen dominiert die normgerechte Zusammenschreibung, gefolgt von der Bindestrichschreibung, die von der Norm zumindest toleriert wird (z.B. *Produktmanager*, *Softwareentwicklungsmethoden*); mehrgliedrige Komposita mit einem Funktionswort als Bestandteil (z.B. *Business-to-Business*) werden ebenfalls meist normkonform mit Bindestrich gebildet und durchgekoppelt, es bestehen hier aber große Unsicherheiten hinsichtlich der Klein-/Großschreibung der einzelnen Bestandteile – nominale Bestandteile müssen hier groß-, nichtnominale Bestandteile kleingeschrieben werden.

Die Autoren beobachten, dass erstens die Anlehnung an den Gebrauch in der Quellsprache (bei den rein englischen Komposita), zweitens die Vertrautheit der einzelnen fremdsprachlichen Elemente und drittens die Länge des Gesamtkompositums eine Rolle bei der Wahl der Schreibweise (getrennt, mit Bindestrich oder zusammen) spielen. Eine Ausrichtung an der Norm dürfte eher zufällig sein, zumal, wie die Autoren im ersten Teil ihrer Arbeit zeigen, sich aus der Norm nur schwer Gebrauchs-Richtlinien ableiten lassen. Dittmann und Zitzke belegen all ihre Befunde mit exakten Zahlen, die sie durch Auszählung der Vorkommen in ihrem Korpus ergeben.

[88] Vgl. Augst (1992), u.A. S. 58.

[89] Vgl. Dittmann und Zitzke (2000), vor allem S. 70-76. Dittmann und Zitzke untersuchen in dieser Hinsicht sowohl die offiziellen Regeln als auch die Richtlinien der Nachrichtenagenturen.

[90] Die Autoren verwenden als Datenbasis die Stellenanzeigen aus der Frankfurter Allgemeinen Zeitung, der Süddeutschen Zeitung und der Welt vom 24. April 1999 und der Neuen Zürcher Zeitung vom 5. Mai 1999. Ihre quantitative Auswertung stützen sie auf die 4225 Vorkommen von Anglizismen in den beiden erstgenannten Zeitungen, vgl. Dittmann und Zitzke (2000), S. 77.

Insgesamt stützt sich die lexikologisch und lexikographisch motivierte Anglizismenforschung sehr stark auf Zeitungskorpora und damit die Pressesprache. Eine Ausnahme bildet die Arbeit, die dem Anglizismenwörterbuch von Carstensen und Busse zugrunde liegt. Die Autoren haben das am IDS verfügbare Freiburger Korpus der gesprochenen Sprache untersucht. Mit der Arbeit von Bartsch liegt zumindest eine varietätenspezifische Untersuchung – und ein entsprechendes Korpus – vor. Integrationsprozesse wurden ebenfalls noch nicht betrachtet. Dies setzt ein systematisches und korpusbasiertes Erfassen von Wortgeschichten voraus. Die entsprechenden Korpora für solche Untersuchungen sind vorhanden, der Aufwand für eine größere Studie, die nicht nur einige wenige Wörter umfasst, ist aber schwer abschätzbar.

7 Partikeln

Eine korpuslinguistisch sehr gute bearbeitete Wortart sind die Partikeln. Wir wollen hier die wichtigsten korpusbezogenen Arbeiten als Beispiele für korpusbasierte linguistische Forschung im Bereich der Wortarten vorstellen.

Es herrscht weitgehend Uneinigkeit darüber, welche Wörter zu den Partikeln zählen und in welche Unterklassen diese Wortklasse zerfällt. Die Duden Grammatik[91] subsumiert die Adverbien, Präpositionen und Konjunktionen unter die Partikeln und wählt damit eine sehr weite Definition, die die meisten nicht flektierenden Wörter umfasst[92]. In einem engeren Sinn verwendet etwa Helbig diesen Begriff[93]. Er bezeichnet mit *Partikel* „solche morphologisch unflektierbaren Wörter, die über keine solchen syntaktischen Funktionen verfügen, wie sie den Wörtern anderer unflektierter Wortklassen zukommen“[94]. Eine noch engere Definition fasst lediglich die Modalpartikeln in diese Kategorie[95]. Helbig unterscheidet die folgenden Subklassen von Partikeln:

- Abtönungs- oder Modalpartikeln (z.B. *auch, bloß, denn*);
- Gradpartikeln (z.B. *auch, gerade, sogar*);
- Steigerungspartikeln (z.B. *außerordentlich, etwas, ganz*);

[91] Vgl. Duden-Grammatik.

[92] Dies stimmt nur so ungefähr, da den Interjektionen ein eigenes Kapitel gewidmet ist.

[93] Vgl. Helbig (1994).

[94] Helbig (1994), S. 20.

[95] Vgl. Helbig (1994), S. 21