Ulrike Haß-Zumkehr

Das Wort in der Korpuslinguistik

Chancen und Probleme empirischer Lexikologie

1. Was ist ein Wort – token, type oder Lemma?
2. Datentypen
3. Korpusaufbau, Korpusorganisation und Recherche-Software
4. Nutzen, Interpretationsbedürftigkeit und -fähigkeit der Analysedaten
5. Literatur

Die Lexikologie versteht sich meist als eine eher theoretisch modellierende Disziplin, die die landläufige Vorstellung, der Wortschatz sei ein kulturell mehr oder weniger wertvolles Sammelsurium von Wörtern, dadurch zu überwinden trachtet, dass sie Relationen und Strukturen zwischen den Elementen des Wortschatzes konstituiert. Angeregt wurde das Systematisieren und Strukturieren durch nichts anderes als lexikografische Arbeit, angefangen bei den Begriffshierarchien der frühen Neuzeit über die wortbildungsbezogenen Systeme der Barockzeit und die Herstellung etymologischer und damit indoeuropäischer Relationen im 19. Jahrhundert bis hin zu den Forderungen Hermann Pauls, die semantischen Vernetzungen des Wortschatzes auch im alphabetisch geordneten Wörterbuch darzustellen (vgl. Haß-Zumkehr 2001).

Doch da Lexikografie in der Regel als bloßes Handwerk, als Redaktionsstätigkeit und „Bearbeiten“¹ bereits vorhandenen Materials angesehen wird – ‚wissenschaftliche Praxis‘ nennt sie Wiegand immerhin –, gerät aus dem Blick, dass alle Systematisierungen der Lexikologie im engeren Sinne ihren Ausgang in eben jener bloßen Praxis nahmen. Dabei spielt keine geringe Rolle, dass dem Wortartikel als dem Endergebnis lexikografischer Arbeit kaum noch anzusehen ist, welche theoretischen Reflexionen und Modellierungen, stets empirisch überprüft an Instanzen der Wortverwendung selbst, der lexikografischen Beschreibung vorangehen mussten. Diese weitgehend unsichtbare Arbeit der lexikografischen „Werkstatt“ (so bezeichnete es vermutlich als erster Daniel Sanders) lässt sich als empirische Lexikologie bezeichnen und soll in diesem Beitrag als eigenständige, d.h. als für die theoretisch orientierte Lexikologie *und* für die lexikografische Praxis relevante Teildisziplin plausibel gemacht werden. Man kann das Dilemma überspitzt formulieren, um das erst noch angemessen wahrzunehmende bzw. zu entwickelnde Zwischenglied deutlich werden zu lassen: Die lexikologischen Modellierungen, insbesondere da diese in ihrer Mehr-

---

¹ Wörterbuchautorinnen und -autoren werden oft als ‚Bearbeiter‘ bezeichnet; damit wird impliziert, dass etwas Vorgängiges (was?) auf eine Weise transformiert wird, die mit der Kreativität und Individualität von ‚Autorschaft‘ nichts gemein hat.

heit übereinzelsprachlich angelegt sind, stehen bisher auf den zu schwachen Füßen von wenigen (zumeist englischen) Beispielen der Individualkompetenz, wohingegen vor allem allgemeinsprachliche Wörterbücher, wenn mit einer systematisch kaum zu bändigenden Flut von Materialien (Belegen) konfrontiert, über positivistische Auflistungen schwer hinausgelangen.

Aber mehr noch als durch die hiermit angedeutete Leerstelle wird die empirische Lexikologie angeregt durch die Möglichkeiten der Korpuslinguistik. Unter Korpus versteht man im weiteren Sinne jede definierte Menge von Texten oder auch von Textausschnitten (Belegen), die die Basis einer linguistischen Untersuchung liefern. Im engeren Sinne ist ein Korpus im Unterschied zum Belegarchiv im Zettelkasten eine Menge elektronisch gespeicherter und digital auszuwertender Texte. Ob man ein Korpus in mehrere Teilkorpora aufgliedert oder einzelne Korpora zu einem Gesamtkorpus zusammenfasst, ist eine Sache des Blickwinkels und auch der technisch-dokumentarischen Organisation. Korpuslinguistik heißt somit, große Textmengen elektronisch zu speichern, gezielt zu durchsuchen und statistisch basierte Analysen zu nutzen. Zur Zeit werden die darin steckenden Möglichkeiten noch selten ausgenutzt; statt dessen werden elektronische Textkorpora meist noch als die schnellere und bequemere Variante des Belegzettels gesehen. Doch so genutzt ist der Aufwand, den der Aufbau von Textkorpora und die Entwicklung von Recherchesoftware speziell für linguistische Zwecke erfordert, nicht zu rechtfertigen. Mit der Bezeichnung Korpuslinguistik werden also technisch-informatische, theoretisch- wie methodisch-linguistische und anwendungsbezogene Aspekte zusammengefasst. Zu den Anwendungsbereichen der Korpuslinguistik (*corpus linguistics*; Lenz 2000) zählen bisher allerdings überwiegend automatische Übersetzung und Spracherkennung (*language engineering*). Zur Beantwortung offener Fragen zu einem einzelsprachlichen System ist Korpuslinguistik bisher noch kaum herangezogen worden (vgl. Teubert 1999, bes. 293). Mit den Chancen für eine empirisch fundierte und theoretisch konsistente Lexikologie, die durch die Korpuslinguistik entstehen, sind sowohl neue Antworten auf bekannte Fragen als auch neue und weiterführende Fragen verbunden.

Nachfolgende Bemerkungen lehnen sich an die öffentlich zur Recherche zugänglichen Korpora des Instituts für Deutsche Sprache an² und damit auch an das institutseigene Recherchesystem Cosmas. Über die genannte Internet-Adresse kann jeder die in diesem Beitrag exemplifizierten Nutzungsarten nachvollziehen.

---

² Zugänglich über Internet-Adresse <http://www.corpora.ids-mannheim.de>. Unter dem Korpusnamen ‚public‘ sind alle aus urheberrechtlichen Gründen freien Texte zusammengefasst. Die übrigen stehen der wissenschaftlichen Nutzung innerhalb des IDS (Mitarbeitern und registrierten Gästen) zur Verfügung.

1. Was ist ein Wort – token, type oder Lemma?

Gegenstand der theoretischen Lexikologie sind einmal das Wort (Lexem), verstanden als bi- oder trilaterales Sprachzeichen, und zum zweiten der Wortschatz, verstanden als System aus Wörtern (Lexemen). Beides wird wesentlich sprechsituationsenthooben gesehen, denn die Lexikologie ist eine langueorientierte Disziplin. Wo situationsgebundene, monosemierte Wörter im wissenschaftlichen Fokus stehen, ist von praktischer Semantik, Pragmatik, Sprachkritik oder sogar von Sprachgeschichte, nicht aber von Lexikologie die Rede.

Für die Korpuslinguistik ist ein Wort dreierlei. Die korpuslinguistische Perspektive nimmt ihren Ausgang von der Zählbarkeit ausdrucksseitig fassbarer Einheiten in elektronisch (v)erfassten Texten geschriebener Sprache.³ Diese Einheiten sind zunächst die Zeichenketten, aus denen sich die Korpus Texte äußerlich zusammensetzen. Je nach Anlage eines Korpus sind dies aber auch Markierungen, die den Korpus Texten im Prozess des Korpusaufbaus hinzugefügt wurden (siehe Abschnitt 3). Die zählbaren Primäreinheiten werden als sogenannte tokens von anderen Sichtweisen des Worts abgehoben, während die Metadaten Informationen zum Kontext jedes einzelnen tokens liefern.

Das Interesse der Lexikologie wie auch der Lexikografie richtet sich allerdings auf systematische Eigenschaften des Wortschatzes, auf das Usuelle und Typische des Wortgebrauchs, nicht auf den Einzelfall der Wortverwendung. Dass Wörterbücher lexikalische Einheiten deshalb aber dekontextualisieren und isolieren, wie Teubert (1999, 304) meint, ist nicht zwangsläufig. Der Grad der Generalisierung, Abstraktion und Typisierung, den eine Bedeutungserläuterung aufweist, hängt wohl vom jeweiligen lexikografischen Konzept ab. Der Unterschied zwischen Typ und Exemplar wird in der terminologischen Unterscheidung zwischen token, type und Lemma wiedergespiegelt. ‚Token‘ oder auch ‚(laufende) Wortform‘ ist jede als lexikalische Einheit infrage kommende Zeichenkette in einem Korpus Text. Die Größe eines Korpus wird in laufenden Wortformen angegeben. Der Jahrgang 2000 der *Berliner Zeitung* umfasst z.B. 22,8 Mio. Wortformen (ohne Werbung). Die Jahrgänge 1994–1999 der Wochenzeitung *Die Zeit* enthalten 24,79 Mio. Wortformen. Das British National Corpus⁴ umfasst ca. 100 Mio. laufender Wortformen. Das Gesamtkorpus des Instituts für Deutsche Sprache in Mannheim (IDS) überschritt im Januar 2001 die Grenze von einer Milliarde (laufender) Wortformen.

Die Zahl der tokens kann also Aufschlüsse über die diachrone Umfangsveränderung einer großen Wochenzeitung, über die durchschnittliche Satzlänge

³ Die Bezeichnung Korpuslinguistik wird durchaus auch mit Bezug auf digitalisierte sprechsprachliche Texte verwendet. Die methodischen Voraussetzungen schon im technischen Bereich sind hier aber so verschieden von denen schriftsprachlicher Korpora, dass gesprochene Sprache im Folgenden ausgeklammert bleiben muss. Zur Korpuslinguistik für gesprochene Sprache siehe <http://www.ids-mannheim.de/prag/links.html> sowie Lenz 2000, 33ff.

⁴ Vgl. <http://info.ox.ac.uk/bnc/>

einzelner Autoren, Textsorten oder Zeitstufen geben, vorausgesetzt es kann ein entsprechendes Teilkorpus, etwa ein Jahrgang, gesondert betrachtet werden. Solche Umfangsquantifizierungen können zusammen mit anderen Faktoren in Schlussfolgerungen einfließen über veränderte Rezeptionsgewohnheiten von Zeitungslesern, die sowohl konzentriertere, kürzere Informationen bevorzugen als auch Text zunehmend im Zusammenhang mit Bildern interpretieren. Auch die durchschnittliche Satzlänge eines (Teil-)Korpus kann im Vergleich aufschlussreich sein: Die Sätze der *Süddeutschen Zeitung* in den Jahren 1995–1999 enthielten durchschnittlich 24,2 tokens, die der *Frankfurter Rundschau* der Jahrgänge 1997–1999 22,3 tokens.

Viele tokens eines Texts sind äußerlich identisch und bilden daher einen type. Ein type ist meist eine bestimmte Schreib-, Flexions- oder Wortbildungsform. Listen von types sind aufschlussreich für die Wortbildungsforschung. Flexionsparadigmen können im Regelfall aus der Sprachkompetenz der Lexikografen ‚richtig‘, d.h. normgerecht gebildet werden. Jedoch zeigt die aus Korpora gewonnene Liste aller types zu einem bestimmten Lexem, dass zwischen der Sprachwirklichkeit und der sowohl morphologischen als auch orthografischen Norm eine größere Differenz existiert, als sie Lexikologen aus ihrer Kompetenz heraus annehmen würden.

Beispiel *online*: Hierzu sind 3136 types, d.h. außer dem Simplex auch Flexionsformen und Zusammensetzungen belegt; von *A-Z-Online-Quellenverzeichnis* bis *Onlineumsatz*. Sieht man sich nur das Simplex *online* an, dann ist das token mit der Schreibung *online* 22.740 mal belegt, das token mit der (Getrennt-)Schreibung *on line* insgesamt nur 353 mal. Dieser enorme Größenunterschied hat zunächst einiges mit dem quantitativen Übergewicht der jeweils jüngsten Jahrgänge im IDS-Korpus zu tun.⁵ Dennoch ist bei einer chronologischen Sortierung der Treffer (tokens) nach Jahrgängen zu erkennen, dass erstens das Lexem überhaupt erst Mitte der 80er Jahre des 20. Jahrhunderts aufkommt und dass zweitens die Getrenntschreibung und die Bindestrichschreibung (*on line*, *on-line*) Mitte der 90er Jahre durch die Zusammenschreibung (*online*) abgelöst worden ist. Ob dies u.a. mit der gesteigerten Produktivität dieses Lexems vor allem innerhalb mehrgliedriger Bindestrich-Zusammensetzungen zusammenhängt, wäre eine Hypothese, die sich durch die Verknüpfung chronologischer mit Frequenzdaten ergibt. Sie dürfte mit Hilfe weiterer Datentypen und Korpusanalysen relativ leicht zu klären sein. Sinnvollerweise macht man solche Untersuchungen aber nicht an einem einzelnen Lexem fest, sondern an einer systematisch definierten Klasse von Lexemen, etwa von Bindestrich-Anglizismen, damit generalisierbare Aussagen gewonnen werden können.

---

⁵ Dies ist aber keine zufällige oder gar von Desorganisation zeugende Besonderheit. Wer je mit der Akquisition von Korpus-texten zu tun hatte, weiß, dass sich Kontakte zu und Bereitwilligkeit von Textgebern im Laufe der Zeit verbessern; quantitativ steigend wirkt sich auch die zunehmende Digitalisierung des Presse- und Buchdrucks aus. Die Quantität von kontinuierlich gepflegten Korpora wächst im Ergebnis exponentiell.

<i>on line/on-line/On line/On-Line</i>		<i>online/Online</i>	
Jahr:	Treffer (token)	Jahr:	Treffer (token)
1985:	2	1985:	2
1986:	4	1986:	2
1987:	2	1989:	3
1989:	2	1993:	152
1993:	7	1994:	220
1994:	5	1995:	1091
1995:	97	1996:	1698
1996:	80	1997:	1953
1997:	54	1998:	2096
1998:	63	1999:	2147
1999:	32	2000:	626
2000:	5		

Bei einem so relativ neuen und entlehnten Wort mag die geschulte Individualkompetenz eine gewisse Vielfalt der Schreibungen vermuten, aber sicher nicht Zeitraum und Verlauf des Wandels vorherzusagen.

Ein anderes Beispiel: Unter den ca. 360.000 acht mal und öfter belegten⁶ Lemmata des gegenwartssprachlichen IDS-Gesamtkorpus befinden sich 315 linkerweiternde Zusammensetzungen auf *-frau* (von *Aboriginesfrau* bis *Zwillingfrau*). Sieht man sich diese Liste genauer an, erkennt man einen neuen Bildungstyp, bei dem stark lexikalisierte und teilweise idiomatisierte Zusammensetzungen auf *-mann* (a) teils aus rechtlichen Gründen ‚moviert‘, (b) teils aber auch aus ironisch-emanzipatorischen Gründen ‚aufgebrochen‘ werden:

- (a) *Außenhandelskauffrau, Bankkauffrau, Fachfrau, Feuerwehrfrau* usw.,
- (b) *Biederfrau, Buhfrau, Eisfrau, Gewährsfrau, Hampelfrau, Jägersfrau, jedefrau, Sauerfrau, Schmerzensfrau, Schneefrau, Seefrau, Wasserfrau, Watschenfrau, Weihnachtsfrau* u.ä.

Die chronologische Sortierung zeigt, dass die in diesem Korpus seit 1985 belegten rechtlichen Bildungen (a) den ironisch-emanzipatorischen (b) um mindestens ein bis drei Jahre vorausgehen. Es zeigt sich ferner, dass *jedefrau* als einziges Lexem dieser Klasse ebenfalls seit 1985 und zudem um mindestens das Dreifache häufiger belegt ist als die anderen Lexeme aus der Klasse (b). Somit fällt es aus (b) heraus. Rückfragen an das Korpus und Interpretation der Ergebnisse fangen an dieser Stelle an, eingehenderes Interesse zu wecken, können hier aber nicht weiter verfolgt werden.

Die Differenz zwischen dem aus der sprachwissenschaftlichen Individualkompetenz Vorhersagbaren und dem Korpusbefund macht nicht nur existierende Normunsicherheiten der Sprachgemeinschaft deutlich (wie im Falle der Schreibung von *online*), sondern auch die Allgegenwart des Normen- und Sprachwandels.

⁶ Die Grenze von acht Treffern zur ersten und raschen Unterscheidung von usuellen von okkasionellen Bildungen ist natürlich willkürlich.

Types lassen sich in der Regel (aber nicht immer) auf eine Grundform zurückführen. Diese Grundform, wie in der Lexikografie *Lemma* genannt, spielt in der Korpuslinguistik, anders als in der Lexikografie, die Rolle des von den Nutzern potenziell eingegebenen Suchworts. Es stellt *keine* anordnungsrelevante Einheit dar. Die Funktion eines korpuslinguistischen Lemmas ist, alle diejenigen types zu bündeln, die Nutzer interessieren könnten, wenn sie ein bestimmtes Suchwort eingeben. Wer *Freiheit* als Suchwort eingibt, möchte auch Belege zu *Freiheiten*, möglicherweise auch zu *Pressefreiheit*, *Freiheitssucher*, *freiheitlich* und unter Umständen auch zu *freyheit*, *freihait*, *friheit* usw. ausgegeben erhalten.

Das Spektrum der Ableitungen und Zusammensetzungen zu einem Lexem, das ein hinreichend großes Korpus dokumentiert, ist beträchtlich größer als das entsprechende Spektrum, das selbst eine Gruppe von Experten ad hoc zusammenstellen könnte. Auch die kompetenzbasierten Annahmen über okkasionelle und usuelle Wortbildungen stimmen mit Sicherheit nicht mit dem überein, was ein Korpus mittels einer anzusetzenden Frequenzgrenze ‚aussagt‘.

Der lexikografische Nutzen solcher type-Listen mag manchem mit wachsender Größe des Korpus als Last erscheinen: Es gilt, die adhoc-Bildungen auszusondern. Dies ist mithilfe von Frequenzkriterien relativ einfach zu erledigen. Setzt man wie oben das Frequenzkriterium z.B. bei acht Treffern an, rutscht allerdings ein type wie *Senkrechtsarg* als usuelle Bildung durch. Bei genauerem Hinsehen befinden sich diese acht Treffer alle in ein und demselben Text, einer Zeitungsmeldung der Art ‚Vermischtes und Kurioses‘, die von einer Erfindung berichtet. Es ist also eine Mindestdistribution nach Texten und am besten auch nach Zeitstufen oder Jahrgängen mit dem Frequenzkriterium zu verknüpfen. Nach einem Aussondern der adhoc-Bildungen bleiben dennoch so viele Wortbildungen übrig, dass die Entscheidung über den Ansatz als Wörterbuch-Lemma schwer fällt, schwerer jedenfalls als die Auswahl aus einer in jedem Falle kleineren, kompetenzbasiert erstellten Liste von Wortbildungen. Hier nun müssen lexikologische Klassifikationen greifen, indem Wortbildungsmuster und die für sie exemplarischen Lexeme gesucht werden. Auch die semantischen Binnenrelationen von Zusammensetzungen liefern klassifikationsrelevante Aspekte, die überdies in kognitiver Sicht interessant sind.⁷

Ein Lemma in der Korpuslinguistik ist zwar vor allem Suchwort, kann aber auch als Vorstufe für das lexikografische Lemma eines potenziellen Wörterbuchs genutzt werden. Eine der grundlegenden Fähigkeiten eines korpuslinguistischen Recherchesystems ist die Zuordnung vieler tokens zu einem type und in einem weiteren Schritt die Abbildung verschiedener types auf ein Lemma. Im Suchvorgang selbst werden diese Schritte in umgekehrter Reihenfolge abgearbeitet. Eine Liste von Lemmata im Hintergrund des Recherchesystems ist Vor-

⁷ Teubert (1999, 297f.) sieht mit Recht eine kaum überbrückbare Kluft zwischen Korpuslinguistik und kognitiver Linguistik. M.E. könnte es aber gelingen, mit Hilfe der Korpuslinguistik zu erweisen, dass kognitive Strukturen von der parole abhängen und sich mit ihr wandeln.

aussetzung dafür, dass zu einem eingegebenen Suchwort zunächst alle relevanten types und dann auch die Belegstellen mit je einem token gefunden werden. Geleistet wird dies von einem sogenannten Lemmatisierer, einer Softwarekomponente,⁸ die einerseits eine Liste von Grundformen enthält, andererseits über Kombinationsregeln verfügt, die ihr erlauben, unbekannte Zeichenketten zu zerlegen und auf bekannte zurückzuführen. Manchmal entstehen dabei Fehler wie *Achillesvers*, *Senftuba*, *Aludosis* und *Drachenfliege*, die nicht nur erheitern, sondern auch einen Eindruck von der Arbeitsweise dieser Softwarekomponente vermitteln. Die Qualität eines Lemmatisierers kann parallel zum Ausbau eines Korpus mit entsprechendem personellen Einsatz kontinuierlich verbessert werden, indem weitere Wortbildungsregeln algorithmisch formuliert und neue Grundformen hinzugefügt werden.

Ein Korpus konfrontiert mit der beträchtlichen Differenz zwischen der Wirklichkeit der einzelsprachlichen Wortbildung und den überlieferungsbedingten Vorstellungen von ‚Grund-‘ bzw. ‚Normal-‘Formen. Deutlich wird dies insbesondere bei den Partizipien im Verhältnis zu ‚ihren‘ Infinitiven. Partizipien wie *belebt*, *entfernt*, *gelungen*, *vergangen*, *geboren*, *geschickt*, *geschlossen* sind so viel häufiger und dazu öfter attributiv denn prädikativ belegt als die Infinitivformen, dass man sich aus lexikologischen Gründen fragen muss, ob diese Partizipien nicht als lexikalisiert zu betrachten und im Wörterbuch entsprechend als eigenes Lemma anzusetzen sind.⁹ Eine semantisch-pragmatische Analyse der morphologisch verwandten types fördert u. U. auch inhaltsseitige Differenzen zutage, die für das Vorliegen unterschiedlicher mentaler Konzepte und gegen eine gemeinsame Behandlung im Wortartikel sprechen.

Einige partizipial gebildete Adjektive bzw. Adverbien scheinen unter Lexikografen als solche anerkannt und sind beispielsweise im Duden-Universalwörterbuch als Lemma angesetzt, z.B. *führend*, *entsprechend*, *laufend*, *anschließend*, *bestimmt*, *geben*, *betont*, *erklärt*, *betroffen*. Anderen, bei denen der Verdacht auf Lexikalisierung ebenso besteht (z.B. *versucht*, *vergangen*, *verkauft*, *bestätigt*, *besetzt*, *geblieben*, *geplant*) wird dieser Status jedoch nicht zugewiesen. Ein korpuslinguistisches Problem dabei ist die Homonymie von Infinitiven und 1. und 3. Person Plural einerseits und 3. Person Singular und Partizip II andererseits. Dort allerdings, wo nicht-homonyme Partizipien I und II häufiger sind als mit diesen homonyme Infinitive und finite Verbformen zusammen genommen, stellt die Frequenz ein starkes Argument für die Hypothese der Lexikalisierung dar. Man vergleiche folgende Frequenzen:

entsprechend	98.483	entsprechen	8.911
anschließend	43.634	anschließen	3.340
betroffen	47.663	betreffen	4.501

⁸ Siehe: <http://www.ids-mannheim.de/kt/merkmal2.html>

⁹ In der Lernerlexikografie wird ohnehin, aber aus anderen Gründen, die gesonderte Lemmatisierung vor allem unregelmäßiger Stammformen des Verbs gefordert bzw. realisiert. Es geht mir hier aber allein um die lexikologischen Begründungen für lexikalisierte und nichtlexikalisierte Wortformen.

vergangen	127.672	vergehen	2.803
laufend	47.564	laufen	39.424

Nur nach weitergehenden Analysen – hier hilft die Feststellung der Kookkurrenzpartner (s.u.) weiter – kann im Fall regelmäßiger Verben die Homonymie von finiter Verbform und (attributiv gebrauchtem) Partizip II aufgelöst werden. Die Frequenzangaben in der nachfolgenden linken Spalte enthalten also beide. Dennoch wird die relative Seltenheit des Infinitivs sehr deutlich:

bestimmt	87.384	bestimmen	13.167
betont	83.344	betonen	5.429
erklärt	238.015	erklären	28.116
verkauft	54.954	verkaufen	27.898
versucht	86.611	versuchen	34.005
bestätigt	57.036	bestätigen	12.167
besetzt	32.817	besetzen	6.054
geplant	119.439	planen	10.480

Ein paralleles Problem ist die Homonymie von Infinitiv und Partizip II (z.B. *beraten*, *bekommen*, *geraten*, *vergessen*, *verlassen*, *erfahren*). Homonymie als ausdrucksseitige Ununterscheidbarkeit von types verhindert aussagekräftige Frequenzdaten, es sei denn die Frequenzanalysen werden mit disambiguierenden Kookkurrenzanalysen kombiniert. Außer dem Frequenzkriterium kommt das der Lexikalisierung (semantischen Selbstständigkeit) in Betracht, wenn über die Behandlung von Partizipien in Wörterbüchern – als Lemma mit eigenem Artikel oder als dem lemmatisierten Infinitiv subsummierter type – zu entscheiden ist. Genau genommen ist die Frequenz nur als ein Indiz für Lexikalisierung zu werten, aber eben doch als ein ‚objektives‘ Indiz. Die gängige Praxis etwa des Duden-Universalwörterbuchs¹⁰ zeigt, dass die Frage nach der semantischen Selbstständigkeit einer ‚lemmafähigen‘ Form intersubjektiv uneinheitlich beantwortet wird, falls sie denn systematisch gestellt wurde. Und das Frequenzkriterium stand mangels Korpus vermutlich nicht zur Verfügung.

Was für das Verhältnis zwischen Infinitiv und Partizipien gilt, gilt ähnlich für das Verhältnis zwischen Singular- und Pluralformen von Substantiven. Nicht nur der Infinitiv, sondern in sehr vielen Fällen gehören auch die Formen des Singulars zu den selteneren Formen im Flexionsparadigma eines Lemmas:

Angabe:	Sg.: 2.647	Pl.: 104.360
Aktie:	Sg.: 21.630	Pl.: 43.631
Einzelheit:	Sg.: 172	Pl.: 7.904
Ermittlung:	Sg.: 2.407	Pl.: 22.035
Geld:	Sg.: 9.386	Pl.: 15.182
Maßnahme:	Sg.: 8.003	Pl.: 25.740
Nachricht:	Sg.: 12.198	Pl.: 16.862
Problem:	Sg.: 72.934	Pl.: 104.421

¹⁰ Hier heißt es in einigen Fällen innerhalb der für etymologische Angaben reservierten eckigen Klammer: „eigentlich Partizip 2 von ...“, in anderen (*geboren*, *bewegt*, d.h. nicht nur bei unregelmäßigen Verben!) wird auf den Infinitiv verwiesen.

Struktur:	Sg.: 13.164	Pl.: 17.831
Verhandlung:	Sg.: 7.462	Pl.: 45.791
Kondition:	Sg.: 2.659	Pl.: 4.484
Datum:	Sg.: 8.365	Pl.: 34.508
aber		
Aktion:	Sg.: 41.812	Pl.: 16.530
Aussicht:	Sg.: 18.381	Pl.: 9.383
Chance:	Sg.: 68.022	Pl.: 50.425
Technik:	Sg.: 46.262	Pl.: 6.464
Technologie:	Sg.: 15.040	Pl.: 6.625
Zeit:	Sg.: 350.932	Pl.: 60.408

Die erste Gruppe zeichnet sich durch den z.T. um ein Vielfaches häufigeren Plural aus. In der zweiten ist zwar der Singular häufiger, aber auch hier besteht der Verdacht auf eine größere oder kleinere semantisch-pragmatische Differenz zwischen Singular und Plural. Vielleicht sollte man neben den bekannten Pluraliatantum, die über keinen Singular verfügen (*Eltern, Leute, Kosten*), eine weitere Klasse usueller (unechter) Pluraliatantum ansetzen, deren Spezifikum nicht auf der morphologischen, sondern auf der semantisch-pragmatischen Ebene liegt. Prototypisch sind hier etwa *Datum* und *Daten*, aber auch *Kondition* und *Konditionen*. Vergleichende Untersuchungen des Gebrauchs von Singular- und Pluralformen können die generellen Gründe für eine (wachsende?) Usualität bestimmter Pluralformen in bestimmten Textsorten feststellen.

Die Frage nach einer möglichen Lexikalisierung stellt sich bei den usuellen Pluraliatantum nicht mit der gleichen Brisanz wie bei den partizipialen Adjektiven, wird aber ebenfalls entscheidend davon abhängen, ob diese Frage korpus- und damit in irgendeiner Form gebrauchsbasiert oder ausschließlich kompetenzbasiert entschieden wird. Die traditionell motivierte Fixierung auf die Normalform Singular verhindert jedenfalls, dass semantisch-pragmatische Besonderheiten des Plurals in den Blick kommen.

Den Infinitiv und den¹¹ Nominativ Singular als die ‚prototypische‘ Form zur Repräsentation mehr oder weniger komplexer morphologischer Paradigmen anzusetzen, war eine Entscheidung der abendländischen Tradition. Zu fragen ist, inwieweit genau diese beiden durch die Tradition hervorgehobenen Formen wirklich prototypisch sind und von den Sprachteilhabern/Nutzern auch entsprechend eingestuft werden. In gedruckten Wörterbüchern ist diese Frage relativ irrelevant, denn es sind die Lexikografen, die die Form des Lemmas vorgeben und darüber entscheiden, ob schwer zuzuordnende Varianten eventuell als Verweislemma angesetzt werden. Wer wie Fremdsprachler nur einen unvollständigen oder unsicheren Überblick über das jeweilige Flexionsparadigma hat und überdies orthografisch unsicher ist, findet das Gesuchte unter Umständen nicht, obwohl es ‚drin‘ steht. In elektronischen Nachschlagewerken, die neben einer Stichwortsuche oft auch über eine Volltextsuche und manchmal auch über

¹¹ In Wahrheit gibt es hier mehrere Möglichkeiten, abhängig von der Kombination mit bestimmtem, unbestimmtem oder gar keinem Artikel: (ein) *Angestellter*, (der) *Angestellte*, (die) *Angestellte*.

eine sogenannte schreibungstolerante Suche verfügen, werden selbst unsichere und in der Tradition der ‚Normalformen‘ ungeschulte Nutzer hingegen immer fündig, wenn das Gesuchte überhaupt im System enthalten ist.

Untersuchungen über die Form, in der Suchwörter (in elektronischen Wörterbüchern und Internet-Suchmaschinen) eingegeben werden, könnten Aufschluss darüber geben, ob bestimmte und wenn ja, welche der types eines morphologisch-orthografischen Paradigmas einen prototypischen Rang haben. Eben die Gewöhnung vor allem von Internet-Vielnutzern an schreibungstolerante Suchmöglichkeiten könnte allerdings dazu führen, dass in der Wahrnehmung der Sprachteilhaber überhaupt gar keine prototypisch ausgezeichneten Formen je Flexionsparadigma mehr existieren, sondern dass man sich von unterschiedlichen orthografischen und morphologischen Varianten her dem Paradigma nähert. Das Variantenspektrum würde dann nicht (mehr) als Paradigma, sondern als äußerlich vielgestaltige Repräsentation eines einheitlichen semantischen Konzepts wahrgenommen.

2. Datentypen

Im ersten Abschnitt dieses Beitrags ist zur Veranschaulichung des lexikologischen Nutzens der Korpuslinguistik nur die Frequenzanalyse herangezogen worden. Sie ist zwar die Basis jeder korpuslinguistischen Lexikologie, aber, wie sich schon andeutete, keineswegs deren Grenze. Lexikologen, für die Programme zur Korpusrecherche mehr oder weniger eine black box sind, können sich je nach Anfrage an das Korpus mit einer ganzen Reihe unterschiedlicher Datentypen versorgen lassen. Diese seien im Hinblick auf ihr lexikologisches Potenzial kurz skizziert:

Die Auflistung der in einem Korpus belegten types klärt nicht nur über morphologisch-orthografische Varianten eines Lemmas, sondern auch über belegte Zusammensetzungen und Ableitungen und deren Usualitätsgrad (über Häufigkeit) auf. Es ist offensichtlich, dass hierdurch Vergleichsuntersuchungen zur Produktivität bestimmter Wortbildungsmuster – synchron nach Textsorten und Varietäten, diachron innerhalb der gleichen Textsorte und Varietät – in einer Weise ermöglicht werden, mit der eine Exzerption per Hand nicht konkurrieren kann.

Der erfahrungsgemäß beste zweite Schritt, nach frequenzannotierten Listen von types, ist die Auseinandersetzung mit automatisch erzeugten Listen signifikanter Kookkurrenzpartner wahlweise zu einem infrage stehenden Lemma oder type. Nicht alle Rechercheprogramme verfügen wie das im IDS entwickelte über eine Kookkurrenz- oder Kollokationskomponente.¹² Aber sie ist entscheidend für die linguistische Qualität der Korpusanalysen, denn differenzierte und

---

¹² Damit ist bei Korpusrecherchen in der Regel dasselbe gemeint. Näheres zum Kollokationsanalyse-Tool des IDS siehe Belica/Steyer (2001).

mit flexiblen Parametern¹³ versehene Kookkurrenzanalysen lassen sich für eine ganze Reihe lexikologischer Fragestellungen verwenden. Sie basieren auf der Korrelation der relativen Häufigkeit einzelner types und Lemmata, auf dem signifikanten Miteinander-Vorkommen zweier oder mehrerer types oder Lemmata innerhalb eines wählbaren Wortabstands. So werden lexikalische Kohäsionen mess- und vergleichbar.

Zu *Hausfrau* sind beispielsweise folgende signifikante Kookkurrenzpartner nachweisbar; ihre Reihenfolge entspricht der absteigenden lexikalischen Kohäsion der Partner; die kursivierten Lexeme verweisen auf weitere Elemente eines Kookkurrenzclusters. In Anschluss daran werden die Kookkurrenzpartner zu einem polysemen Lexem aufgelistet und danach kommentiert:

Kollokatoren zu *Hausfrau*.¹⁴

BelegNr	Gamma Kollokatoren	Häufigkeit
1+614:	3446 Mutter	614
615+69:	835 männer	69
684+115:	695 Rentner	115
799+36:	450 Hausmänner	36
835+30:	426 Berufsorganisation <i>Mozartstr</i>	30
865+3:	<i>Berufsorganisation</i>	3
868+82:	364 Studenten	82
950+60:	242 Mütter	60
1010+19:	237 Berufstätige	19
1029+40:	229 Blume	40
1069+31:	222 Kindern	31
1100+42:	217 geb <i>geboren</i>	42
1142+20:	<i>geb</i>	20
1162+28:	211 Hausmann	28
1190+20:	178 Selbständige	20
1210+28:	156 Angestellte	28
1238+18:	155 Kantonsrätin	18
1256+11:	143 geboren	11
1267+20:	138 Sekretärin	20
1287+9:	131 Hausmännern	9
1296+20:	122 Krankenschwester	20

¹³ Dem mit dem Programm Unvertrauten sagen die dort genannten Parameter wohl wenig. Dennoch sei als ein Eindruck von der Komplexität qualitativ hochwertiger Kollokationsanalyse wiedergegeben, worunter linguistische Nutzer wählen können: Lemmatisierung ja/nein („Sollen Lemmata oder Textwörter gesucht werden?“); Wortabstand: 1 bis n; „Soll über Satzgrenzen hinaus gesucht werden?“ ja/nein; Granularität: fein/mittel/grob („Wie intensiv sollen Mehrwortgruppen gesucht werden?“); Zuverlässigkeit hoch/normal/analytisch („Ziehen Sie Ausbeute oder Zuverlässigkeit vor?“); Clusterzuordnung: eindeutig/mehrfach („Was tun, wenn ein Beleg mehreren Kollokationsclustern zugeordnet werden kann?“); Funktionswörter beachten: ja/nein; Autofokus ja/nein („typische Stellung der Kollokatoren im Kontext ermitteln“); Anzeigen des sog. Gammawerts ja/nein (er gibt die Stärke der lexikalischen Kohäsion an). In Belica/Steyer (2001) werden die Parameter erstmals beschrieben.

¹⁴ Kollokatoren sind sortiert absteigend nach der berechneten Stärke der lexikalischen Kohäsion, angegeben durch den ‚Gammawert‘. Die kursiv gesetzten Lexeme markieren Clusterzugehörigkeit.

1316+20:	121 Arbeitslose	20
1336+2:	118 Berufsfrau	2
1338+13:	117 Rentnern	13
1351+20:	117 verheiratet	20
1371+10:	107 berufstätige	10
1381+11:	105 tüchtige	11
1392+29:	104 Beruf	29
1421+19:	99 Lehrerin	19
1440+14:	97 brave	14
1454+10:	91 Bäuerin	10
1464+1:	89 kaufmännische	1
1465+7:	88 züchtige	7
1472+7:	88 Wurmrieder	7
1479+11:	83 gelernte	11
1490+13:	78 Einkaufen	13
1503+15:	78 Handwerker	15
1518+10:	76 verheiratete	10
1528+9:	74 biedere	9
1537+57:	73 Kinder	57
1594+18:	71 Schüler	18
1612+10:	70 biederer	10
1622+27:	69 Küche	27
1649+5:	66 männern	5
1654+7:	65 Kauffrau	7
1661+8:	60 Ferres	8
1669+5:	60 gestreßte	5
1674+14:	59 Wäsche	14
1688+18:	58 Arbeiter	18
1706+3:	56 Familienväter	3
1709+13:	56 Kochen	13
1722+17:	56 Maria	17
1739+24:	56 Rolle	24
1763+7:	55 Einkaufstaschen	7
1770+12:	55 Müttern	12
1782+5:	54 Nänni	5
1787+6:	52 Erwerbstätige	6
1793+3:	52 Gemeinderätin	3
1796+7:	51 frustrierten	7
1803+5:	51 Köchin	5
1808+52:	51 alte	52
1860+2:	51 gard	2
1862+2:	51 Kindergärtnerin	2
1864+4:	51 Mutterdasein	4
1868+5:	50 Rentnerinnen	5
1873+9:	49 Erika	9
1882+8:	49 Landwirt	8
1890+9:	48 Herisau	9
1899+5:	48 Geschäftsleute	5
1904+2:	48 Geschäftsfrau	2
1906+11:	48 Elisabeth	11
1917+4:	47 Haushilfen	4
1921+4:	47 Blankers-Koen	4
1925+5:	47 Bäuerinnen	5
1930+8:	46 perfekte	8
1938+4:	46 Teilzeitangestellte	4

1942+2:	44	Veronika	2
1944+7:	44	60jährige	7
1951+15:	44	Ehemann	15
1966+2:	43	Maaßen	2
1968+5:	43	berufstätig	5
1973+3:	42	Ehefrau	3
1976+3:	42	Psychiatrieschwester	3
1979+5:	42	Bichwil	5
1984+4:	42	Bügeln	4
1988+18:	41	Monika	18
2006+5:	41	Mutterrolle	5
2011+5:	41	Hobbyköche	5
2016+3:	41	Familienfrau	3
2019+4:	41	Marianne	4
2023+4:	40	Magdalene	4
2027+2:	40	Romi	2
2029+3751:		<i>statistisch unspezifisch</i>	3751

Kollokatoren zu *Absatz*:¹⁵

BelegNr	Gamma	Kollokatoren	Häufigkeit
1+328:	4929	reissend finden	328
329+83:		<i>reissend</i>	83
412+1:	4010	Artikel Grundgesetz 2 3 1	1
413+14:		Artikel Grundgesetz 2 3	14
427+8:		Artikel Grundgesetz 2 1	8
435+60:		Artikel Grundgesetz 2	60
495+9:		Artikel Grundgesetz 3 1	9
504+48:		Artikel Grundgesetz 3	48
552+47:		Artikel Grundgesetz 1	47
599+36:		Artikel Grundgesetz	36
635+2:		Artikel 2 3 1	2
637+16:		Artikel 2 3	16
653+23:		Artikel 2 1	23
676+111:		Artikel 2	111
787+12:		Artikel 3 1	12
799+71:		Artikel 3	71
870+69:		Artikel 1	69
939+133:		Artikel	133
1072+117:	2834	Paragraph 1	117
1189+257:		Paragraph	257
1446+111:	1923	reissenden finden	111
1557+32:		<i>reissenden</i>	32
1589+6:	1371	Grundgesetz	6
1595+267:	1198	finden	267
1862+215:	948	steigern	215
2077+339:	781	steigen	339
2416+515:	695	Prozent	515
2931+169:	596	Umsatz	169
3100+143:	550	Produktion	143

¹⁵ Kollokatoren sind sortiert absteigend nach der berechneten Stärke der lexikalischen Kohäsion, angegeben mittels ‚Gammawert‘. Die kursiv gesetzten Lexeme markieren Clusterzugehörigkeit.

3243+287:	530	2	287
3530+22:	527	Satz 1	22
3552+35:		Satz	35
3587+72:	505	ankurbeln	72
3659+67:	417	stagnieren	67
3726+123:	391	Produkt	123
3849+102:	356	sinken	102
3951+87:	344	Schuh	87
4038+134:	320	3	134
4172+301:	282	hoch	301
4473+159:	262	1	159
4632+3:	236	Einkommensteuergesetz	3
4635+68:	220	weltweit	68
4703+1:	198	2c	1
4704+7:	190	Paragraf	7
4711+30:	188	schleppend	30
4741+37:	173	rückläufig	37
4778+51:	163	Fahrzeug	51
4829+30:	159	Halbjahr	30
4859+23:	159	Heizöl	23
4882+23:	156	Inland	23
4905+13:	154	alkoholfrei	13
4918+56:	153	4	56
4974+65:	151	fördern	65
5039+19:	146	Nutzfahrzeug	19
5058+11:	129	Preisgarantie	11
5069+26:	128	schrumpfen	26
5095+19:	122	gestiegen	19
5114+6:	116	Mineralölprodukt	6
5120+1:	115	Nr	1
5121+50:	115	erhöhen	50
5171+19:	114	Geschäftsjahr	19
5190+18:	110	Erzeugnis	18
5208+11:	108	harzen	11
5219+9:	106	kehrtmachen	9
5228+15:	103	Audi	15
5243+22:	101	Stiefel	22
5265+23:	100	Pkw	23
5288+9:	97	mengenmässig	9
5297+9:	96	Getränk	9
5306+27:	93	Marke	27
5333+19:	91	stocken	19
5352+9:	87	Rover	9
5361+2:	86	Bundesverfassung	2
5363+6:	85	Umsatzrekord	6
5369+15:	84	quartal	15
5384+3:	84	BMW	3
5387+3:	83	Ertrag	3
5390+9:	83	Personalcomputer	9
5399+7:	83	Ertrag	7
5406+14:	77	klettern	14
5420+3:	76	überprüfungsbedürftig	3
5423+2:	75	landwirtschaftlich	2
5425+4:	74	Stöckelschuh	4
5429+13:	70	regeln	13

5442+19:	68 Steigerung	19
5461+2:	68 zweistellig	2
5463+8:	zweistellig	8
5471+7:	68 Sohle	7
5478+15:	65 flach	15
5493+3:	65 Grundgesetzartikel	3
5496+1:	65 flach	1
5497+2:	63 Ausländergesetz	2
5499+15:	60 Rindfleisch	15
5514+1:	60 wachsen	1
5515+10:	60 verdoppeln	10
5525+37:	60 wachsen	37
5562+12:	60 verdoppeln	12
5574+1:	60 wachsen	1
5575+2:	59 Spezialität	2
5577+5:	59 Skoda	5
5582+34:	58 Tonne	34
5616+6:	58 Umsatzzahl	6
5622+1:	58 besagen	1
5623+3:	58 Benz	3
5626+1:	58 besagen	1
5627+6:	58 Benz	6
5633+11:	57 Kilowattstunde	11
5644+2:	57 Satzung	2
5646+23:	57 Vorjahr	23
5669+12:	57 schleppen	12
5681+6:	57 Gesamtjahr	6
5687+1:	57 Vorjahr	1
5688+44:	55 sichern	44
5732+5:	55 Hektolitern	5
5737+1:	55 sichern	1
5738+2:	55 Lire	2
5740+6:	54 Nordamerika	6
5746+4:	54 Reissender	4
5750+1:	52 EG	1
5751+6:	52 Neuwagen	6
5757+3:	51 Autohersteller	3
5760+8:	49 hergestellt	8
5768+4:	48 Beschaffungsmarkt	4
5772+6:	47 Westeuropa	6
5778+8:	47 Hektoliter	8
5786+51:	47 Stück	51
5837+15:	46 Zeile	15
5852+28:	46 1995	28
5880+5:	45 Zentimeter	5
5885+11:	45 boomen	11
5896+17:	44 zurückgehen	17
5913+3:	43 Alpinski	3
5916+18:	43 leicht	18
5934+1:	42 Ausland	1
5935+2:	42 anfügen	2
5937+18:	42 Ausland	18
5955+1:	42 Dieselmotortreibstoff	1
5956+13:	42 mangeln	13
5969+1:	42 Dieselmotortreibstoff	1

5970+7:	41 spitz	7
5977+3:	41 Dataquest	3
5980+1:	41 spitz	1
5981+1:	40 Ottokraftstoff	1
5982+1:	40 inländisch	1
5983+4:	40 Mineralbrunnen	4
5987+4:	40 inländisch	4
5991+4:	inländisch	4
5995+2:	40 Ottokraftstoff	2
5997+4001:	statistisch unspezifisch	4001

Es wird unmittelbar deutlich, welche unterschiedlichen Gründe die hier nachgewiesenen Kohäsionen haben können: Es gibt mehr oder weniger variable oder feste Syntagmen, ob man sie den Phraseologismen, Funktionsverbgefügen (*reißenden Absatz finden*), Kollokationen (*züchtige, biedere Hausfrau; reißender, schleppender Absatz*), oder irgendeiner anderen Klasse zuordnen will. Ferner gibt es pragmatisch bzw. diskursindizierte und sachgebietsspezifische Kohäsionen (zu *Absatz*: *Artikel, Paragraf, Satz, Grundgesetzartikel, Einkommenssteuergesetz; steigern, Umsatz, Produktion, stagnieren, boomen; Schuh, Stöckelschuh, Stiefel, Sohle, flach, spitz*).¹⁶

Mit dem Instrument der Kookkurrenzanalyse erhält man folglich Zugang zu Hypothesen über usuelle (!)¹⁷ Mehrworteinheiten, über das Spektrum der lexikalischen wie syntaktischen Variabilität einer Mehrworteinheit, sowie über unterschiedliche Lesarten eines Lexems, insbesondere, wenn diese parallel zu Domänen und Themen angesetzt werden sollen. Die an entsprechenden Kookkurrenzpartnern abzulesende Polysemie von *Absatz* z.B. ist nach dem hier zugrundegelegten Textkorpus dreifach:

- (a) Verkauf,
- (b) Teil eines Textes, insbesondere eines Gesetzestextes,
- (c) Teil eines Schuhs.

Auch die *Reihenfolge* lässt sich aus dem Korpus begründen: Die überragende Häufigkeit der wirtschaftsbezogenen Lesart (a) sorgt für Prototypikalität. Die textbezogene Lesart (b) ist relativ häufig, aber das Lexem wird hier rein referenziell zur Identifizierung von Textstellen und völlig beiläufig im Zusammenhang von Diskussionen über aktuelle Rechtsfragen verwendet. Am seltensten ist noch die schuhbezogene Lesart (c). Die Reihenfolge a, b, c ist im Zehnbandigen Duden (2000) genau umgekehrt wie hier. Es fällt ferner auf, dass eine weitere, in Wörterbüchern regelmäßig auftauchende Lesart hier fehlt: die architekturbezogene, von fachsprachlichen Lesarten zu schweigen:

¹⁶ Zur Differenzierung s. Steyer 2000.

¹⁷ Das Ausrufungszeichen bezieht sich auf die immer wieder festzustellende Diskrepanz zwischen den in einem weitgehend kompetenzgestützt erarbeiteten Wörterbuch verzeichneten Mehrworteinheiten und den durch Usualität (Korpusfrequenz) ausgezeichneten Mehrworteinheiten.

4. *Unterbrechung einer Fläche*, von etw. Fortlaufendem: der A. eines Berges, einer Treppe, einer Mauer. 5. <Pl. selten> (Geol.) *Ablagerung*: A. von Schlamm und Kies. [© 2000 Dudenverlag]

Man kann daraus schließen: Die Duden-Lesart (4) von *Absatz* ist so selten, dass sie in einem überwiegend gebrauchts- und pressesprachlichen Korpus des gegenwärtigen Deutsch nicht signifikant häufig belegt ist. Möglicherweise ist sie als fachspezifisch einzustufen. Dass die Duden-Lesart (5) die einzige fachspezifische Lesart von *Absatz* ist, scheint zweifelhaft; sie kann daher in einem standardsprachlichen Wörterbuch auch entfallen.

Wenn man auch betonen muss, dass automatische Kookkurrenzanalysen nichts weiter als Hypothesen zu unterschiedlichen syntagmatischen und semantischen Aspekten liefern – die linguistische Beurteilung der jeweiligen Kohäsion wird ein Computer m.E. auch in alle Zukunft nicht leisten können –, so ist der Wert solcher Hypothesen doch offensichtlich.

Für eine Validierung der Hypothesen und für weitere Fragestellungen eignet sich der Datentyp der sogenannten Key-Word-in-Context-Belege (KWICs), bei dem ein strikt zeichenbegrenzter, maximal eine Bildschirmzeile langer Satzausschnitt rechts und links des types, Lemmas oder auch der zuvor festgestellten signifikanten Kookkurrenz angezeigt wird. KWIC-Beispiele zu *Absatz*:

WKB/SG3, *Der Spiegel* (2. Hj. 1990)

wei Senatoren hat - bestimmt Artikel 51 *Absatz* 2 des Grundgesetzes : .

die Schuhfabrik hat keinen *Absatz* mehr , beim Glühlampenhersteller

Verteidigungswaffen fänden " reißenden *Absatz* " , berichtet das Hallesche Boule

WKB/ZT2, *Die Zeit* (1. Hj. 1990)

fkleber fanden unter Trabifahrern guten *Absatz*.

lzog sich im Schatten des Artikels 20, *Absatz* 2 der Verfassung , der die Diskri

WKD/BZA, *Berliner Zeitung*, 2. Halbjahr 1989

r Forschung über die Produktion bis zum *Absatz* - weltmarktfähige Erzeugnisse her
ndung von Wissenschaft , Produktion und *Absatz* in diesen starken ökonomischen Ei

KWIC-Beispiele zu *Absatz* mit Kollokator *stieg*:

M89 us 17,7 Prozent). Dagegen stieg der *Absatz* bei BMW of North America Inc (ebe

M89 passen: In wenigen Jahren stieg der *Absatz* von alkoholfreiem Bier von prakti

M91 (gfu) in Hannover hingewiesen. Der *Absatz* stieg auf 7,9 (1990: 6,9

M94 6 000 Wagen. In Brasilien stieg der *Absatz* um 22,4 auf 344 000 Fahrzeuge. In

M95 runnen in Bonn mitteilte, stieg der *Absatz* in den ersten drei Quartalen 1995

M96 zent gegenüber 1994. Zwar stieg der *Absatz* im Dezember 1995 bei einer Jahres

M98 von 15 Prozent. Im Januar stieg der *Absatz* von Mercedes- Pkw im Vergleich zu

H85 den drei folgenden Jahren stieg der *Absatz* auf jeweils etwa 900 Mill. DM. gr

MK1 so-Ausstellung machte selbst diesen *Absatz* ansehnlich und originell - stieg

BZK n. als im Laufe des Jahres 1968 der *Absatz* wieder stieg , haben sich in den

C93 rie ihre Umsätze um 30 Prozent. Der *Absatz* von Windows-Anwendungen stieg
sog

U95 r erfolgreichen C-Klasse, stieg der *Absatz* um etwa 15 Prozent auf mehr als 5

Mittels KWIC-Indizes kann nicht nur die polyseme Struktur des Lexems sowie seine Rolle in Mehrwortausdrücken und die Qualität einer lexikalischen Kohäsion festgestellt, sondern auch viele der syntaktischen Merkmale ermittelt werden. Im Falle des oben genannten Beispiels *online* ist dies etwa die Nachstel-

lung bei attributiver Funktion, die Verwendung in Zusammenhang vor allem mit einigen bestimmten Verben (*gehen*) und die sehr viel häufigere Verwendung als präfixartiges Element (vgl. Haß-Zumkehr 2000). Dadurch, dass jede KWIC-Zeile mit einem Quellennachweis (Sigle) versehen ist, erfüllt dieser Datentyp bereits die Bedingungen des wissenschaftlichen, nachweisenden Belegs.

Der letzte hier zu nennende Datentyp sind die altbekannten Belege, die ein Lexem oder eine Kollokation im größeren Satz- bzw. Textkontext zeigen. Doch weisen Belege aus elektronischen Korpora Unterschiede gegenüber handexzerpierten Belegen in den klassischen Zettelarchiven auf. Die „Korpuslinguistik [wertet] das Korpus systematisch und nicht nur exemplarisch aus“ (Teubert 1999, 302). Der Datentyp des Belegs kann je nach Rechtersystem mehrere Subtypen aufweisen: Die Belegstellenliste zeigt in Form dokumentierter Abkürzungen, wie viele Belege in einem Subkorpus und in einem Text enthalten sind. Dies genügt, um eine an einen einzelnen Text gebundene *adhoc*-Bildung wie das oben erwähnte *Senkrechtsarg* trotz höherer Frequenz zu identifizieren:

Auszug aus der Belegstellenliste zu *Absatz*:

1+13:M89/901	<i>MM</i> Januar 1989	13
14+10:M89/902	<i>MM</i> Februar 1989	10
24+25:M89/903	<i>MM</i> März 1989	25
49+19:M89/904	<i>MM</i> April 1989	19
68+30:M89/905	<i>MM</i> Mai 1989	30
98+33:M89/906	<i>MM</i> Juni 1989	33
131+35:M89/907	<i>MM</i> Juli 1989	35
166+25:M89/908	<i>MM</i> August 1989	25
191+21:M89/909	<i>MM</i> September 1989	21
1463+2:WKB/MM3	<i>Mannheimer Morgen</i> (2. Hj. 1990)	2
1465+10:WKB/RM2	<i>Rheinischer Merkur</i> (1. Hj. 1990)	10
1475+1:WKB/RM3	<i>Rheinischer Merkur</i> (2. Hj. 1990)	1
1476+4:WKB/SG3	<i>Der Spiegel</i> (2. Hj. 1990)	4
1480+1:WKB/ST2	<i>stern</i> (1. Hj. 1990)	1
1481+2:WKB/TZ1	<i>taz</i> , Sonderheft 1 (2. Hj. 1989)	2
1483+1:WKB/ZT1	<i>Die Zeit</i> (2. Hj. 1989)	1
1484+3:WKB/ZT2	<i>Die Zeit</i> (1. Hj. 1990)	3
1487+3:WKD/BZA	<i>Berliner Zeitung</i> , 2. Halbjahr 1	3
1490+1:WKD/BZB	<i>Berliner Zeitung</i> , 1. Halbjahr 1	1
1491+6:WKD/BZC	<i>Berliner Zeitung</i> , 2. Halbjahr 1	6
1497+1:WKD/ND1	<i>Neues Deutschland</i> , [Tageszeitung]	1

Einen wichtigen Aspekt stellt die Verbindung dieser Listen mit den KWIC-Zeilen und den daran anschließenden Belegen dar. Von den Belegstellen kann man durch Mausklick unmittelbar zu einer KWIC-Übersicht gelangen und von diesen zu den Kontextbelegen. Insofern dient die KWIC-Zeilen-Übersicht auch der gezielteren Auswahl von Belegen, nachfolgend solche zu *Absatz* mit Kollokator *stieg* mit je einem Satz vor und nach dem Satz, in dem die beiden Lexeme vorkommen:

In der Oberklasse meldete die Daimler-Vertriebstochter Mercedes-Benz of America Inc (Montvale/New Jersey) für die ersten sechs Monate einen Absatzrückgang von 44 066 im 1. Halbjahr 1988 auf 36 265 (minus 17,7 Prozent). Dagegen *stieg* der *Absatz* bei BMW of

North America Inc (ebenfalls Montvale) auf 34 849 (34 787) leicht an. Das Schlußlicht bildete die Porsche Cars North America Inc, Reno (Nevada), die mit 4556 (8806) nur halb so viel Wagen verkaufte. (M89/907.24363: *Mannheimer Morgen*, 08.07.1989, *Wirtschaft; Absatz deutscher Autos in USA schwächer*)

Bei Rover war der Land-Rover das Zugpferd. Der Absatz des Geländewagens stieg um 20 Prozent auf den Rekordwert von 116 000 Stück, alles in allem erhöhte der Rover-Absatz jedoch nur um 2 Prozent auf 483 000 Wagen. Besserung erwartet sich BMW auch auf dem deutschen Markt, von neuen Modellen und einem verbesserten Vertrieb. (M96/604.15230: *Mannheimer Morgen*, 03.04.1996, *Wirtschaft; Zwei Töchter lagen BMW auf der Tasche*)

Der Umfang der Belege sollte prinzipiell frei wählbar sein, bewegt sich aus urheberrechtlichen Gründen aber in den meisten Systemen zwischen dem Minimum eines Satzes und dem Maximum einer Bildschirmseite.¹⁸ Im Korpusrecherchesystem des IDS können symmetrische und asymmetrische Kontextmengen gewählt werden, etwa ein Satz vor und drei Sätze nach oder drei Absätze vor und kein Absatz nach dem type bzw. Lemma. Im Falle der Hypothese etwa eines Neologismus, einer fachsprachlichen Bedeutung und der Suche in einem fachsprachlichen Subkorpus kann es sinnvoller sein, den Fokus auf den nachfolgenden Kontext zu legen, weil bei Ersterwähnungen von Fachlexemen in einem Text die terminologischen Erläuterungen unmittelbar folgen (vgl. Teubert 1998b). Geht es z.B. um die möglichst scharf abgrenzende Feststellung des Referenzbereichs eines Kollektivums oder um die Kohyponyme zu einem infrage stehenden Hyperonym, dürfte der vorangehende Kontext aufschlussreicher sein.

Solche methodischen Überlegungen bedürfen aber noch einer systematischen Überprüfung, die die thematische Entfaltung in Texten eines bestimmten Typs je spezifisch mit den semantischen und pragmatischen Eigenschaften eines Lexems oder einer Lexemklasse korreliert. Ergebnis wären nicht nur eine effektivere Handhabung des Korpus durch Lexikografen, sondern – im Sinne der empirischen Lexikologie – Erkenntnisse über den Zusammenhang von Text und Diskurs, letzteres hier verstanden als intertextuelle Ebene, und den systematischen Eigenschaften des Wortschatzes. Wenn nämlich paradigmatische Relationen eines Lexems wie Synonymie, Hyperonymie usw. korpusbasiert ermittelt werden sollen, müssen sich diese Relationen irgendwie aus der Ebene des Textes, aus seiner semantischen Struktur ableiten lassen.

Die abgestufte Folge und Verknüpfbarkeit von Belegstellenlisten, KWIC-Zeilen und Belegen ist mehr als nur komfortabel. Bei großen Korpora kann auf diese Weise die unüberschaubare Menge von Belegen gezielt und weil schrittweise reduziert werden. Die einzige Alternative hierzu ist eine methodisch kaum zu begrüßende Zufallsauswahl. Kleinere Korpora sind keine Alternative.

¹⁸ Potenzielle Textgeber wollen verhindert sehen, dass ein kompletter Text über die Belegausgabe ‚abgesaugt‘ werden kann.

3. Korpusaufbau, Korpusorganisation und Recherche-Software

Die Untersuchung des Worts mit korpuslinguistischen Mitteln bindet Lexikologie und Textlinguistik zusammen. Umso entscheidender ist die Zusammenstellung des jeweils zugrundegelegten Korpus. Die Erfahrungen der letzten Jahrzehnte haben gezeigt, dass weder ein für allemal feststehendes kanonisches Korpus wie das British National Corpus (URL: <http://info.ox.ac.uk/bnc/>) noch ein gegenüber textlinguistischen Aspekten indifferentes Korpus wie das des Leipziger Informatik-Projekts Deutscher Wortschatz (<http://www.wort-schatz.uni-leipzig.de>) oder wie die meisten computerlinguistisch orientierten Korpus-sammlungen (z.B. des IMS Stuttgart: <http://www.ims.uni-stuttgart.de/projekte/corplex/>) die Anforderungen sowohl syntaktischer als auch lexiko-logischer Interessen abdecken. Ein breites Anforderungsspektrum abzudecken ist aber angesichts des hohen Aufwands, den der Aufbau eines größeren Korpus kostet, unerlässlich. Von ausschließlich morphologisch-syntaktisch interessierten Computerlinguisten wird Korpuslinguistik oft zu eng definiert und Anforderungen lexikologischer Art tendenziell als „traditionell“ und „philologisch“ abgewertet. Die verstärkte Aufmerksamkeit, die die kognitive Linguistik seit einiger Zeit dem ‚Lexikon‘ als mentaler Komponente gegenüber walten lässt, klammert aber gerade alle pragmatischen und textlinguistischen Aspekte aus ihrem Semantikbegriff aus.

Ohne die Diskussion hier im Einzelnen zu referieren (vgl. Sinclair 1998; unter historischem Aspekt hierzu Bergmann 1988), sei aus der jahrzehntelangen Erfahrung des IDS mit Textakquisition und Korpusaufbau in Form einiger Thesen zusammengefasst, was ein für lexikologische Zwecke geeignetes Korpus ausmacht:

- Die Alternative zu einem abgeschlossenen Korpus ist die Parallelität eines erweiterbaren, aktuell wachsenden (sog. opportunistischen) Korpus mit einem aus dessen Teilen je nach Forschungsziel systematisch komponierten und somit definierten (sog. virtuellen) Korpus. Aus dem sich ständig verändernden opportunistischen Korpus können immer wieder neue, aktualisierte oder nach sonstigen Kriterien gebildete, etwa projektspezifische Korpora gebildet werden. Eine Teilmenge des IDS-Korpus „W-PUB“ (öffentlich zugängliche Texte geschriebener Sprache) ist beispielsweise das virtuelle Korpus WK (Wendekorpora West und Ost), seinerseits untergliedert in Westtexte und Osttexte sowie in Phasen.¹⁹ Ein opportunistisches Korpus ist so groß und so breit gestreut wie möglich ohne Einschränkung durch qualitative Kriterien. Die schiere Größe (in tokens) steigert die Qualität und Feinheit der so wichtigen Kookkurrenzanalysen. Überhaupt sind frequenzbasierte Analysen nur auf der Basis größerer Korpora seriös.

¹⁹ Letztere Untergliederung wird erst mit der neuen Version des Recherchesystems Cosmas getrennt durchsuchbar sein.

- Ein repräsentatives Korpus im eigentlichen, d.h. statistischen Sinne kann es nicht geben, weil die sog. Grundgesamtheit, auf die Repräsentativität bezogen und definiert wird, nicht abgegrenzt werden kann (vgl. Haß-Zumkehr 1998, 71, nach Rieger 1979; Teubert 1998b, 147ff.). Was ist ‚die‘ deutsche Sprache, ‚die‘ Standard-, Gemein-, Gegenwarts-, Umgangssprache, ‚die‘ Fachsprache X? So unterschiedlich und immer nur vage diese Fragen beantwortet werden können, so unterschiedlich kann die Frage nach der Zusammenstellung eines entsprechenden Korpus beantwortet werden. Besser ist es umgekehrt: Man hat ein virtuelles Korpus mit einer bestimmten Zusammensetzung sowie mit einer bestimmbareren Relevanz für die Sprachwissenschaft und begrenzt die Gültigkeit der daraus resultierenden (lexikologischen) Erkenntnisse auf den jeweiligen Sprachbereich des Korpus. Im Übrigen können weder die klassischen Zettelarchive noch Individualkompetenz Aussagen von größerer Generalisierbarkeit machen als die, die auf der Basis eines wie auch immer zusammengesetzten Korpus entstehen.
- Das Korpus muss aus vollständigen Texten bestehen, nicht aus einzelnen Sätzen bzw. Belegen und auch nicht aus willkürlich gekürzten Textausschnitten. Die Texte sollten textlinguistisch annotiert sein.
- Die Texte müssen im Prozess des Korpusaufbaus mit Markierungen, bibliografischer Art (zu Autor, Erscheinungsdatum usw.), textstruktureller Art (zu Überschriften, Absatzgrenzen, Zitaten, Bildern, Tabellen usw.), pragmatischer Art (zu Textfunktion, Sach- oder Fachbereich, Region, Situation usw.) und auch syntaktischer Art versehen werden, um diese Ebenen der Sprache mit der des Wortschatzes verknüpfen zu können (vgl. Haß-Zumkehr 1998, bes. 80).²⁰ Die Metadaten werden den Korpustexten bisher noch überwiegend händisch, teilweise automatisch hinzugefügt. Die händischen Verfahren sind sehr aufwändig; die automatischen vor allem im Bereich der Syntax noch recht fehleranfällig. Hier muss also je nach Forschungsziel und personellen wie finanziellen Ressourcen abgewogen werden. In der Folge ist ein völlig standardisiertes Annotationsschema für alle Korpora, selbst wenn man nur an die zur deutschen Gegenwartssprache denkt, illusorisch. Erreicht wird jedoch eine gewisse Einheitlichkeit der textstrukturellen Markierungen, soweit diese aus den typografischen Eigenschaften der Texte erschlossen werden können. Als internationale Standards haben sich in diesem Zusammenhang die Annotationsschemata der Text Encoding Initiative (TEI) und der europäische Corpus Encoding Standard (CES) etabliert. Es ist jedoch zu beobachten, dass zumindest die Korpus-‚Bauer‘ im deutschsprachigen Bereich diese Standards in der Regel modifizieren, damit vor allem die Metadaten den jeweiligen Forschungszielen angepasst werden.

---

²⁰ Die in den Text mittels Sonderzeichen eingefügten Markierungen sind in der Belegausgabe überwiegend nicht zu sehen, weil sie den Text unübersichtlicher machen. Möglich wäre aber ein wahlweises Hin- und Herschalten zwischen beiden Ansichten. Wichtig ist, dass die Markierungen zu Quelle, Datierung usw. vom Recherchesystem eindeutig identifiziert werden können.

- Ein forschungspraktisches Hindernis muss einkalkuliert werden, das, wenn zu spät bemerkt, zum echten Problem werden kann: die wachsenden Vorbehalte der Rechteinhaber (Autoren, deren Erben oder Verlage) gegenüber einer digital vervielfältigenden Nutzung ihrer Texte. Nicht ganz zu Unrecht wächst die Befürchtung vor ‚Datenklau‘, Raubkopie oder kommerzieller Ausnutzung in dem Maße, wie Internet, e-commerce, e-book und dergleichen an Raum gewinnen. Auch eine erwiesenermaßen nicht-kommerzielle Nutzung der Texte durch wissenschaftliche Institutionen muss mehr oder weniger einschränkenden Bedingungen der Textgeber (z. B. Verbot der online-Nutzung, Verbot der Weitergabe an andere Forscher, Begrenzung des Belegkontextes) folgen.
- Recherchewerkzeuge für Korpora werden an vielen wissenschaftlichen Institutionen, aber auch von Firmen kommerziell entwickelt und man kann sie zum Teil auch erwerben.²¹ Manche sind einzelsprachunabhängig, andere -spezifisch. Letzteres dürfte für semantische und lexikologische Untersuchungen wichtiger sein. Der bei der Forschungsförderung bevorzugt behandelte Bereich der automatischen Übersetzung greift hingegen auf einzelsprachunabhängige Werkzeuge zurück. Ein für das jeweilige Forschungsziel optimales Werkzeug wird in jedem Fall einen großen Anteil an eigener Arbeit bei der Entwicklung und Pflege von Recherchesoftware erfordern – dies ist aus verschiedenen Gründen keine leichte Aufgabe für Universitäten sowie für außeruniversitäre Forschungseinrichtungen. Das ‚schönste‘ Korpus ist aber nur so gut wie das Instrument, das es erst erschließt. Zu den Qualitätsanforderungen – allgemeiner Bedienungskomfort, Schnelligkeit, Sicherheit einmal vorausgesetzt – an ein lexikologisch brauchbares Recherchewerkzeug gehören nicht nur ein ‚gutes‘ Wortsegmentierungsprogramm (tokenizer), eine aktuelle bzw. der jeweiligen Zeitstufe angepasste Grundformenliste sowie angemessen algorithmisierte Wortbildungsregeln (letzteres beides machen einen guten Lemmatisierer aus), sondern auch die Kombinationsuche und die flexibel parametrisierbare Kookkurrenzanalyse.

Der Hinweis auf wissenschaftspraktische Probleme (Urheberrecht, Software) soll keineswegs abschrecken, sondern im Gegenteil die Etablierung einer korpusbasierten Lexikologie von blauäugigem Idealismus entlasten, der letztlich nur zu Resignation führen kann.

---

²¹ Wegen der raschen Veränderungen auf diesem Gebiet verzichte ich auf konkrete Angaben. Am besten informiert man sich per Internet-Suche (Suchwörter „corpus“ und „tools“). Vgl. Lenz 2000, 28ff.

4. Nutzen, Interpretationsbedürftigkeit und -fähigkeit der Analysedaten

Der Computer ist kein Automat, der, nachdem er mit Texten gefüttert wurde, Erkenntnisse über Wörter und Wortschatz auswirft. Korpuslinguistische Verfahren versorgen Lexikologen und Lexikografen mit teilweise neuen Datentypen, die wissenschaftlich zu interpretieren sind. Einige interpretierende Schlussfolgerungen etwa aus Listen von *types* und von *Kollokationen* sind in diesem Beitrag schon gezogen worden. Weiteres ist noch zu benennen.

Die o.g. Liste der 2487 linkerweiternden Substantivkomposita (*Lemmata*) auf *-frau* lässt sich über die kritisch-ironischen Transformationen lexikalisierter Maskulina hinaus noch weiter semantisch klassifizieren und mit einer entsprechenden Klassifikation der noch bedeutend zahlreicheren *movierten* Personenbezeichnungen auf *-in* vergleichen.²²

---

²² *Aspirin* u.ä. nicht movierungsbedingte Bildungen auf *-in* sind natürlich ausgesondert worden. Ich danke meinem Kollegen Ulrich Schnörch für die Versorgung mit korpusbasierten Listen auf *-in*, *-frau* und *-euse*.

Bildungen X+frau	movierte Bildungen X+in
Berufsbezogen (lexikalisiert/nicht lexikalisiert) (<i>Zugehfrau, Putzfrau, Anchorfrau, Bürokauffrau</i>)	berufsbezogen (nicht lexikalisiert) (<i>Lehrerin</i>); berufsbezogene Lexikalisierungen sind selten: <i>Sekretärin</i> ist nicht die ‚weibl. Form von Sekretär‘
Funktionsbezogen (politische Funktion) (<i>CDU-Frau</i>)	Funktionsbezogen (politische Funktion) (<i>Parlamentarierin</i>)
funktionsbezogen (kirchlich-religiöse Funktion/Amt) (<i>Ordensfrau</i>)	Funktionsbezogen (kirchlich-religiöse Funktion/Amt) (<i>Priesterin</i>)
funktionsbezogen (gesellschaftlich) (<i>Geschäftsfrau, Alibifrau</i>)	auf ökonomische Rolle bezogen (<i>Verbraucherin</i>)
sportbezogen (<i>Handballfrau</i>)	sportbezogen (<i>Handballerin</i>)
auf Ehemann bezogen (<i>Zahnarztfrau</i>)	[<i>Bäckerin</i> im Sinne von ‚Frau des Bäckers‘ ist heute veraltet]
auf Ethnie bezogen (<i>Zulufräü, Eskimofrau</i>)	nationalitätsbezogen (<i>Amerikanerin</i>)
	auf Ort und Region bezogen (<i>Berlinerin, Schwäbin</i>)
auf die äußere Gestalt/Aussehen bezogen (<i>Zweizentnerfrau</i>)	bildungsbezogen (<i>Abiturientin</i>)
imagebezogen (<i>Businessfrau, Zweitfrau, Dutzendfrau</i>)	einstellungsbezogen (<i>Darwinistin</i>)
	auf beliebige Tätigkeiten/Rollen bezogen (<i>Gratulantin</i>)

Da Movierungen auf -in sich zum kritisch-ironischen Aufbrechen maskuliner Personenbezeichnungen nicht eignen, ist das Muster des Typs *Hampelfrau* nicht übertragbar. Den sehr vereinzelt vorkommenden ungrammatischen Bildungen *Geistin, Mitgliedin, Mitgliederin, Ritterin* (ungrammatisch?), *Rüdin, Kätzin* fehlt das Ironisch-Augenzwinkernde. Sie sind eher bewusst feministisch motiviert und folgen im Übrigen historischen und literarischen Vorbildern (*Gutsherin, Kaufmännin, Rätin*).

Die leitende Frage einer mit der obigen Tabelle veranschaulichten vergleichenden Klassifizierung könnte sein: Welche semantischen bzw. pragmatischen Aspekte werden bei der Bildung weiblicher Personenbezeichnungen fokussiert, im Unterschied etwa zu männlichen? Wie sind semantische Gemeinsamkeiten (*Handballfrau ~ Handballerin*) und Unterschiede (*Arztfrau/Ärztin*) zu erklären? Warum sind manche Klassen nur einseitig besetzt oder spezifisch akzentuiert? Welche der Einseitigkeiten gehen auf das Konto strukturell-morphologischer Restriktionen (**Eskimoin*) und welche sind eher pragmatisch bedingt? Inwiefern unterscheidet sich die Semantik des Simplex *Frau* von der des Elements -*frau*? Die Interpretierbarkeit allein solcher Listen hat sich in den genannten Fragen noch nicht erschöpft. Zur Beantwortung dieser und weiterer Fragen sind

gezielt Belege heranzuziehen. Ausschließlich anhand großer Belegmengen wäre man auf die Fragen allerdings kaum gekommen.

Ein wesentliches Moment der wissenschaftlichen Interpretation jedes Datentyps ist die Kenntnis über die Zusammensetzung des benutzten (virtuellen, s.o.) Korpus. Inbezug auf die oben erwähnten kritisch-ironischen Bildungen wie *Biederfrau*, *Buhfrau*, *Gewährsfrau*, *Hampelfrau*, *jedefrau*, *Watschenfrau*, *Weihnachtsfrau* könnte man vermuten, dass sie ausschließlich oder ganz überwiegend in bekannt kreativen Presseorganen wie *taz* und *Spiegel* bezeugt sind. Entsprechend zu relativieren wäre das Gewicht dieses Bildungsmusters in der Entwicklung der Standard- oder Gemeinsprache. Die Überprüfung der Belegung zeigt indes ein ganz anderes Bild. Alle genannten Lexeme werden in einem größeren Spektrum von deutschen und schweizerischen Tages- und Wochenzeitungen verwendet, die nicht gerade zu sprachlichen Neuschöpfungen neigen: *Süddeutsche Zeitung*, *Frankfurter Rundschau*, *Mannheimer Morgen*, *St. Galler Tageblatt*, *Zürcher Tagesanzeiger*; *Spiegel* und *taz* sind gerade *nicht* darunter. Die Belege datieren in der Mehrzahl seit 1989, vereinzelt seit Mitte der 90er Jahre. Textbelege wurden für dieses Bild übrigens noch gar nicht benötigt. Deren methodische Funktion konzentriert sich in der Korpuslinguistik zunehmend auf die semantisch-pragmatische Feinarbeit.

Ein weiteres Moment, das bei der Interpretation vor allem von Frequenzdaten beachtet werden muss, ist die Tatsache, dass aktuelle Themen der öffentlichen Diskussion lexikalisch markiert und bestimmte Lexeme deshalb – zeitlich oft vorübergehend – häufiger sind, prototypisches Beispiel war der *Elchtest*. Frequenz muss also in Relation zur Zeit und zur ‚Karriere‘ eines Themas gesetzt werden – das heißt: Frequenz steht in Zusammenhang mit Diskurs (vgl. Teubert 1999). Dabei sind es keineswegs immer schlüsselwortähnliche Lexeme wie *Entsorgung*, *Schwarzgeld*, *Berliner Republik*, die ein Thema begleiten oder sogar indizieren. Das Beispiel von *Absatz* für ‚Textteil eines Gesetzes‘ hat gezeigt, dass diese Funktion auch semantisch-pragmatisch ‚uninteressanten‘ Wörtern zukommen kann; *Absatz* in dieser Bedeutung begleitete gleich mehrere rechtspolitische Themen der letzten Jahrzehnte.

Unter den 1000 häufigsten Wörtern des IDS-Gesamtkorpus sind auffallend viele Spezifikatoren für Quantoren wie *Prozent*, *Mark*, *Seite*, *Jahr*, *Stück*, *Tag*, *Grad*; auch die Frequenz der genannten Lesart von *Absatz* lässt sich aus dieser Perspektive interpretieren. Sie spielen ihre Rolle bei der von der Pressebeurichterstattung allgemein erwarteten Genauigkeit und Präzision. Genauigkeit bei Quantitäten ist viel leichter zu erreichen als Genauigkeit bei Qualitäten. Zu ‚Fakten‘ gehören also wesentlich ‚Zahlen‘, und zu Zahlen gehören Bezeichnungen für das, was jeweils gezählt wird.

Korpuslinguistisch gewonnene Daten bedürfen nicht nur einer fachlinguistischen bzw. lexikologischen Interpretation, sie sind, wie gezeigt werden sollte, auch vielfältig interpretierbar. Die unterschiedlichen Datentypen sind aufeinander bezogen und erlauben in ihrer Gesamtheit neue Einblicke in die Struktur des Wortschatzes. Die Frage allerdings, was denn ‚das‘ bzw. ein lexikalisches System sei, wie viel es mit dem mentalen Lexikon und wie viel mit den Wörtern in

Texten zu tun hat, muss die theoretische Lexikologie, die sich den Möglichkeiten der Korpuslinguistik öffnet, wieder einmal neu beantworten.

5. Literatur

- Belica, Cyril: Statistische Analyse von Zeitstrukturen in Korpora. In: Teubert, Wolfgang (Hrsg.): *Neologie und Korpus*. Tübingen 1998, 31–42. (Studien zur deutschen Sprache, 11).
- Belica, Cyril/Kathrin Steyer: Die COSMAS-Kollokationsanalyse – statistisches Modell, Funktionsweise und Interpretationsspielräume. 2001 (in Vorbereitung).
- Bergmann, Rolf: Probleme der Textauswahl für einen elektronischen Thesaurus. Beiträge zum ersten Göttinger Arbeitsgespräch zur Historischen Deutschen Wortforschung 1. und 2. November 1996. Hrsg. im Auftrag der Kommission für das deutsche Wörterbuch von Jacob Grimm und Wilhelm Grimm von Rolf Bergmann. Stuttgart usw. 1998
- Duden-Universalwörterbuch. 3. Aufl. CD-Rom-Version 1.1. Mannheim 1996.
- Duden: Das große Wörterbuch der deutschen Sprache. 10 Bde auf CD-Rom. Mannheim 2000.
- Feldweg, Helmut/Erhard W. Hinrichs (Hrsg.): *Lexikon und Text. Wiederverwendbare Methoden und Ressourcen zur linguistischen Erschließung des Deutschen*. Tübingen 1996 (Lexicographica Serie Maior, 73).
- Haß-Zumkehr, Ulrike: Das historische Korpus des Instituts für deutsche Sprache. In: Bergmann, Rolf: Probleme der Textauswahl für einen elektronischen Thesaurus. Beiträge zum ersten Göttinger Arbeitsgespräch zur Historischen Deutschen Wortforschung 1. und 2. November 1996. Hrsg. im Auftrag der Kommission für das deutsche Wörterbuch von Jacob Grimm und Wilhelm Grimm von Rolf Bergmann. Stuttgart usw. 1998, 69–80.
- Wortschatz ist mehr als „viele Wörter“. Die Aufgaben der Abteilung Lexik des IDS. In: *Sprachreport* 2/2000, 2–7.
  - Deutsche Wörterbücher – Brennpunkt von Sprach- und Kulturgeschichte. Berlin 2001.
- Journal for Corpus Linguistics. Ed by Wolfgang Teubert (John Benjamins publishing company). Vol. 1, Nr. 1/1996 – Vol. 5, Nr. 1/2000.
- Lenz, Susanne: *Korpuslinguistik*. Tübingen 2000 (Studienbibliographien Sprachwissenschaft, 32).
- Sinclair, John: *Korpusmorphologie. Ein Klassifikationsrahmen*. In: Teubert, Wolfgang (Hrsg.): *Neologie und Korpus*. Tübingen 1998, 111–128. (Studien zur deutschen Sprache, 11).
- Steyer, Kathrin: *Usuelle Wortverbindungen des Deutschen. Linguistisches Konzept und lexikografische Möglichkeiten*. In: *Deutsche Sprache* 28, 2000, 101–125.
- Die korpusbasierte Kookkurrenzanalyse – ein empirischer Weg zur Verifizierung usueller Wortverbindungen. (unveröff. Vortragsmsskpt. für Tagung „Phraseologie und Parömiologie – neue Perspektiven der Forschung“, 17.–22. Juni 2001, Ascona, Schweiz) 2001.
- Teubert, Wolfgang (Hrsg.): *Neologie und Korpus*. Tübingen 1998 (Studien zur deutschen Sprache, 11). [= Teubert 1998a]
- Teubert, Wolfgang: *Korpus und Neologie*. In: Teubert, Wolfgang (Hrsg.): *Neologie und Korpus*. Tübingen 1998, 129–170. (Studien zur deutschen Sprache, 11). [= Teubert 1998b]
- *Korpuslinguistik und Lexikographie*. In: *Deutsche Sprache* 27, 1999, 292–313.