Erschienen in: Kratochvílová, Iva/Wolf, Norbert Richard (Hrsg.): Grundlagen einer sprachwissenschaftlichen Quellenkunde. Tübingen: Narr, 2013. S. 49-66. (Studien zur Deutschen Sprache 66)

Sabine Krome

Digitale Datenflut: Chancen und Tücken eines Textkorpus zur deutschen Gegenwartssprache

Anforderungsprofil, Methoden und Instrumentarien zur Beobachtung des aktuellen Sprach- und Schreibgebrauchs

'Korpusbasiert' oder 'korpusgeleitet' – das Pro und Contra dieser beiden Analyseansätze ist im wissenschaftlichen Diskurs inzwischen intensiv beleuchtet worden, allerdings keinesfalls mit eindeutigen Ergebnissen.[1] Wie auch immer man die beiden Ansätze bewerten mag[2] – generell richtet sich die Frage, welche Vorzüge welche Methode gegenüber welcher anderen hat, ganz wesentlich nach Art und Beschaffenheit des analysierten Korpus sowie nach dem spezifischen Untersuchungskontext und den Zielen, die mit der Analyse verfolgt werden. Die Erarbeitung und Bearbeitung von Wörterbüchern zur deutschen Gegenwartssprache erfolgt heute (in Kombination beider Methoden) vorrangig 'korpusbasiert' – das Korpus ist also in erster Linie Instanz zur Prüfung, zur Verifizierung oder Falsifizierung bestimmter vorgegebener Fragestellungen. Während in der Lexikographie zunächst frequenzorientierte Analysen das wichtigste Paradigma zur Bewertung darstellten, jedenfalls bei der Erstellung von Wörterbüchern, ist eine solche Methodik heute für qualifizierte Auswertungen längst nicht mehr ausreichend. Es geht nicht mehr vorrangig darum, immer mehr Daten zu sammeln und vorweisen zu können. Vielmehr stellt es mittlerweile eine Herausforderung dar, der digitalen Datenflut mit intelligenten Mitteln zu begegnen bzw. Auswertungskriterien zu entwickeln, die den Strom kanalisieren können. Die Einbeziehung sekundärer und tertiärer Quellen ist dabei unabdingbar.[3] Dies gilt umso mehr, als bei der spezifischen Verwendung eines Korpus in der Wörterbucharbeit immer auch ein Abgleich mit großen bereits vorhandenen Wörterbuch-Substanzen, also anderen 'Korpora', den wichtigsten 'sekundären Quellen', erfolgt. Wie muss ein digitales Korpus beschaffen sein, das hier eine sinnvolle Orientierungsnorm bietet?

---

[1] Vgl. z.B. Klosa (2007), Lemnitzer/Zinsmeister (2010, S. 32-38), Lüdeling (2007) zu den verschiedenen korpusanalytischen Ansätzen und Methoden in der Lexikographie.

[2] Dazu etwa die gegenteiligen Positionen von Wolf (2010, S. 20) und Mindt (2010).

[3] Vgl. dazu im Detail Klosa (2007, 2010). Zudem „sind die Korpora [egal in welcher Größe] doch auf jeden Fall endlich, und das heißt, im Hinblick auf das zu Modellierende in zufälliger Weise unvollständig" (Eichinger 2010, S. 27).

Drei Aspekte sind bei heutiger lexikographischer Arbeit an allgemeinsprachlichen Wörterbüchern von zentraler Bedeutung:

1) die Erweiterung der Stichwortsubstanz durch Neologismen,
2) die Aktualisierung der vorhandenen Substanz sowie
3) die Schreibbeobachtung vor dem Hintergrund der Arbeit des Rates für deutsche Rechtschreibung.

Vor diesem Hintergrund soll es im vorliegenden Beitrag um Faktoren gehen, die beim Aufbau des *WAHRIG Textkorpus*^{digital} eine Rolle spielten und die bei der Korpuspflege wesentlich sind, um die textlichen Grundlagen des Korpus sowie um die Methoden und Instrumentarien, die für die Aktualisierung und Weiterentwicklung von Wörterbüchern und für eine qualifizierte Beobachtung des Schreibusus in Zukunft entwickelt werden sollten. Dies impliziert die Frage, was ein solches Korpus, das zum großen Teil Zeitungskorpus ist, leisten kann und was nicht. Und das führt schließlich zu der Überlegung, wie das Korpus mittel- und langfristig weiter ausgebaut werden könnte.

1. Ein Korpus zur deutschen Gegenwartssprache

Die digitale Datenflut ist in Online-Wörterbüchern kein so schwerwiegendes Problem.[4] Print-Wörterbücher mit dem gebotenen Umfang zwingen jedoch dazu, Prioritäten zu setzen und in besonderer Weise auf die Ausgewogenheit des Korpus zu achten.

Das Ziel und der spezifische Nutzen von allgemeinsprachlichen Wörterbüchern ist es, die deutsche Gegenwartssprache authentisch widerzuspiegeln. Dass ein Korpus dazu einen angemessenen Umfang haben sollte, liegt auf der Hand, es sollte zumindest eine ‘relative Vollständigkeit’ des Wortschatzes erreicht sein. Damit verbunden ist, dass es so repräsentativ wie möglich aufbereitet sein sollte. Dies betrifft sowohl den Wortschatz im Gesamtzusammenhang wie auch die repräsentativen ‘Zielgruppen’ eines Wörterbuchs.[5] Im Idealfall sollte jeder Benutzer jeder relevanten Zielgruppe das finden, was er nachschlägt. Das dritte wichtige Kriterium ist das der Aktualität, denn nur so kann die Gegenwartssprache angemessen beschrieben werden. Das Korpus sollte nur in geringem, möglichst repräsentativem Umfang historische Texte enthalten und unbedingt bis in die Gegenwart reichen.

---

[4] Zur Entwicklung von Online-Wörterbüchern mit Hilfe korpusanalytischer Methoden vgl. Bubenhofer in URL 1.

[5] Vgl. dazu ausführlich Krome (2010).

Mit einem Umfang von mittlerweile mehr als 2 Milliarden Wortbelegen ist gewährleistet, dass alle wichtigen Wörter und Wendungen des deutschen Wortschatzes im *WAHRIG Textkorpus* vertreten sind.[6] Wichtig für den Lexikographen ist, dass ein Korpus, welches den dargestellten Anforderungen entspricht, gut strukturiert, leicht zugänglich und textsortenspezifisch aufbereitet ist, also speziell entwickelt für allgemeinsprachliche Wörterbücher mit ihren Subgenres, z.B. Fremdwörtern, Synonymen etc. Spezialwortschätze sollte es in angemessenem Umfang als Sub- und Teilkorpora darstellen können, z.B. den der gehobenen Stilebene. Und es sollte sich auf Bereiche konzentrieren, die allgemeinsprachlich interessant sind, den Wortschatz von gruppenspezifischen Sprachteilnehmern, etwa Jugendlichen, aber 'repräsentativ' als Jugendsprache abbilden. Dies erfordert eine einheitliche Strukturierung der Daten und eine differenzierte Kodierung der Metadaten – nach Wortart, Ressort, Stilebene, regionalem Vorkommen und anderen Kriterien. Die technische und strukturelle Aufbereitung ist essentiell, wie aber werden die genannten Kriterien vom Text- und Wortmaterial und von der Sortierung her im *WAHRIG Textkorpus* umgesetzt?

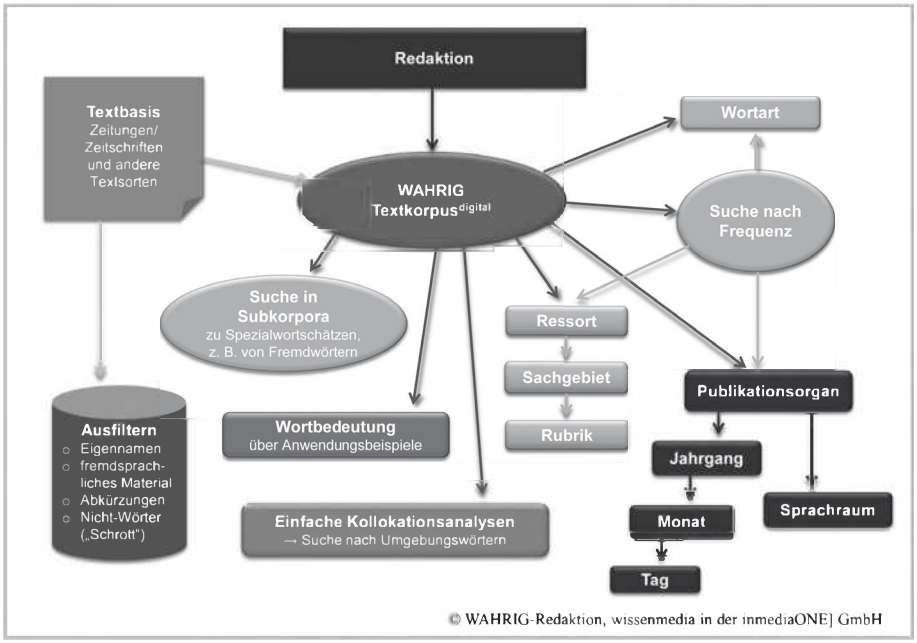


Abb. 1: *WAHRIG Textkorpus*^{digital}: Aufbau und Funktionen

[6] Das *WAHRIG Textkorpus*^{digital} wurde in Kooperation mit der Universität des Saarlandes, Saarbrücken, gezielt für die Arbeit an allgemeinsprachlichen Wörterbüchern entwickelt. Es umfasst Wortbelege u.a. aus folgenden Medien: *Berliner Zeitung*, *Süddeutsche Zeitung*, *Der SPIEGEL*, *Neue Zürcher Zeitung*, *Der Standard*, *Spektrum der Wissenschaft*, *FÜR SIE*, *BRAVO*.

2. Themen- und Sachbereiche, Auswertungskriterien und Instrumentarien

Warum ist ein Zeitungskorpus so interessant? Zeitungen und Zeitschriften sind die auflagenstärksten Publikationen überhaupt. In der überregionalen Ausrichtung großer Zeitungen wird ein breites Publikum erreicht und über die deutsche Standardsprache angesprochen. In ihrer weiten Verbreitung über verschiedenste Ziel- und Altersgruppen decken sie annähernd alle Themenbereiche ab. Dies spiegelt das *WAHRIG* Textkorpus wider: Hier sind alle wesentlichen Themenbereiche, Sachgebiete und Zielgruppen erfasst. Genau dies sind auch die zentralen Bereiche, auf die der Wortschatz der deutschen Gegenwartssprache gegründet ist. Das Thema Politik/Zeitgeschehen spielt im Korpus die größte Rolle, dicht gefolgt vom Bereich Wissenschaft/Technik. Dies sind auch beim Entstehen von Neologismen die produktivsten Bereiche, sie sind damit wichtig für die Aktualisierung des Wortschatzes im Wörterbuch. Das Korpus kann sowohl nach Sachgebieten wie nach Zeitschriften und Jahrgängen durchsucht werden, dabei ist sehr genau festzustellen, welche Information aus welcher Quelle stammt.

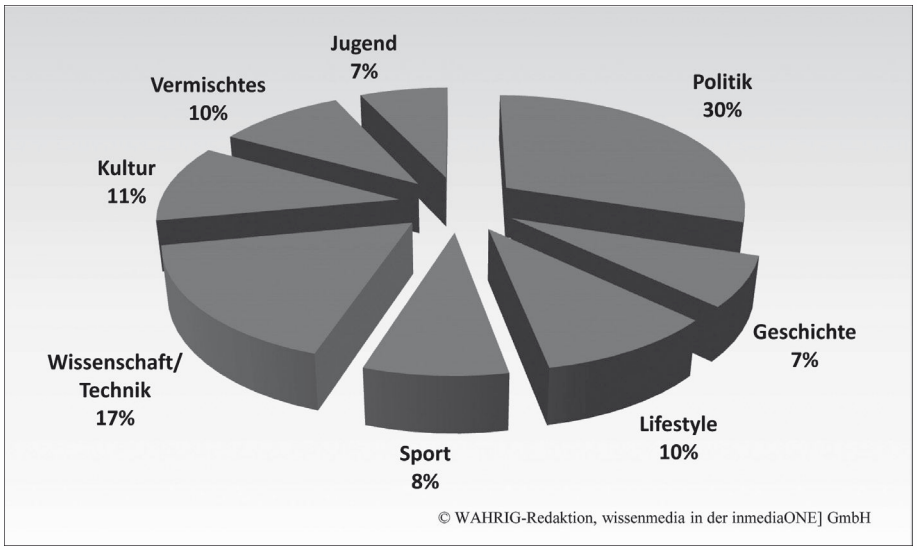


Abb. 2: Sachbereiche im Korpus

Diese Differenzierungsmöglichkeiten machen einmal mehr deutlich, dass eine mechanisch-statistische Frequenzanalyse nicht gewährleistet, dass der deutsche Wortschatz authentisch und damit im weitesten Sinne repräsentativ widerge-

spiegelt wird.[7] Zur Bestimmung der Aktualität allerdings sind Frequenzen essentiell. Neben der absoluten Frequenz über den Gesamtzeitraum der Erfassung hinweg, die häufig sehr problematisch ist und Ergebnisse verfälschen kann,[8] ist die Frequenz im Vergleich der einzelnen Jahrgänge ein zentrales Bewertungskriterium.

3. Aktualität und Neologismen

Zur Widerspiegelung aktueller Tatbestände und Entwicklungen sind Zeitungstexte geradezu prädestiniert. Sie beleuchten tagesaktuell alle wichtigen Themen und Ereignisse, die die Sprachteilnehmer zum gegenwärtigen Zeitpunkt bewegen. Keine andere Textsorte könnte diese Voraussetzungen erfüllen, die sonst nur die gesprochene Sprache und zum Teil die Kommunikation in Internet, z.B. über Blogs und Kommunikationsforen, bietet. Die Daten haben in diesem Rahmen wichtige Funktionen:

1) Im Vergleich verschiedener Jahrgänge sind interessante Neologismen aufzuspüren.
2) Im Gegenzug ist ebenfalls nachweisbar, dass bestimmte Lemmata weniger frequent sind und (allmählich) gar nicht mehr vorkommen (veraltende/veraltete Wörter).
3) Die Anwendungsbeispiele sind empirisch belegt und authentisch. Dadurch wird ein ‘Wiedererkennungseffekt’ beim Wörterbuchbenutzer ausgelöst.

Dies zeigt sich im Kontext einer der wichtigsten Aufgaben der Lexikographie – der Neologismenarbeit. Nach Herberg/Kinne/Steffens (2004) spricht man von Neologismen, wenn etwas Neues entsteht und eine Benennung braucht oder wenn ein Sachverhalt sich neu konstituiert oder darstellt (‘Neulexem’). Ein Neologismus kann ebenso eine Wortneubildung sein, die ältere Begrifflichkeiten ersetzt oder ergänzt. Ein bereits vorhandenes Lexem kann aber auch eine neue Bedeutung erhalten (‘Neubedeutung’).[9]

---

[7] Wie etwa im methodischen Zugriff bei Quasthoff (2007, S. 9) suggeriert. Vgl. dagegen Scherer (2006, S. 49).

[8] Dies ist zu einem großen Teil die Ursache für die Unzulänglichkeit von Ergebnissen, die die Suche über eine öffentlich zugängliche Suchmaschine wie *Google* in einem „opportunistischen Korpus“ (Meger 2010, S. 102), dem Internet, häufig zur Folge hat.

[9] Vgl. Herberg et al. (2004, S. XI). Zum Terminus ‘Neologismus’ und dem Begriff der ‘Usualität’ vgl. auch Meger (2010, S. 13-25), Quasthoff (2007, S. 7-9), Elsen (2011, bes. S. 19-22), Elsen/Dzikowicz (2005, S. 80). Die Definition von ‘Neologismus’ ist in der Forschung nicht klar fixiert. Vgl. dazu Wolf-Bleiß (2009, S. 85f.).

3.1 *Flashmob* und *Schweinegrippe*: Neologismen versus Okkasionalismen

Ein solches ‘Neulexem’, das einen neuen Sachverhalt beschreibt, bezeichnet *Flashmob*. Der Begriff umschreibt eine „spontane Ansammlung von Menschen, die gemeinsam eine überraschende Aktion durchführen, die vorab im Internet verabredet wurde“.[10] Das Wort ist erst seit 2003 im Korpus belegt, es braucht eine gewisse Anlaufzeit, bis es in der Öffentlichkeit registriert wird, und erlebt eine Bedeutungsspezifizierung hin zum Bereich Politik. Eine solche Form der politischen Aktion haben erst die modernen elektronischen Kommunikationsmedien möglich gemacht, das iPhone und das Internet. Entsprechend steigt die Frequenz des Wortes von 2008 bis 2010 um ein Vielfaches.

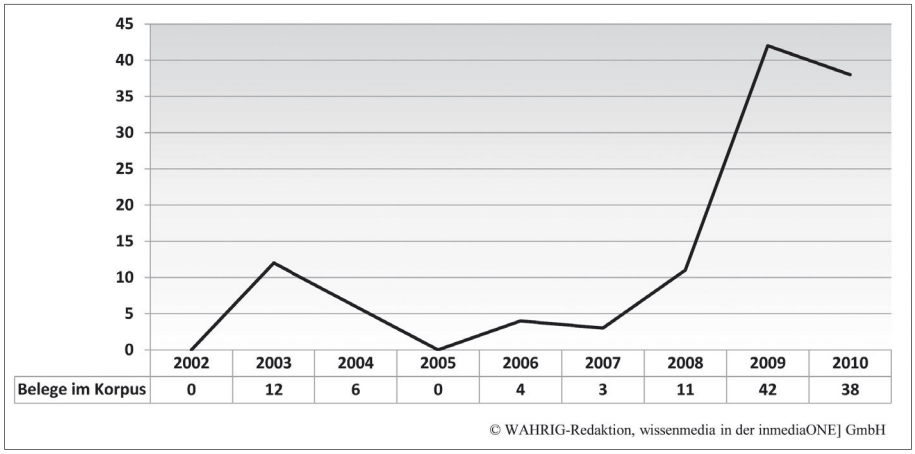


Abb. 3: Neologismen: Korpusanalyse zu *Flashmob*

Die gegenteilige Entwicklung ist bei einem anderen vor allem im Jahr 2009 hochfrequenten Lemma zu beobachten, der *Schweinegrippe*. Ausgehend von 0,2% vor dem Jahr 2009 steigt die Frequenz des Wortes 2009 auf 87,4% des Gesamtvorkommens, um dann bereits 2010 mit dem erfolgreichen Bekämpfen der Seuche wieder auf nur 12,4% zu fallen. Es handelt sich also bei diesem Neologismus bzw. der früher bereits vorhandenen Bezeichnung für eine Krankheit um eine Eintagsfliege.[11] In anderen Textsorten als aktuellen Zei-

[10] Brockhaus *WAHRIG Deutsches Wörterbuch* (2011).

[11] Lemnitzer/Zinsmeister fassen den Begriff ‘Okkasionalismus’ sehr eng als Gelegenheitsbildung, die nicht lexikalisiert ist (vgl. Lemnitzer/Zinsmeister 2010, S. 147), breiter im Vergleich Elsen (2011, S. 21). In ein aktuelles allgemeinsprachliches Wörterbuch würden solche Lemmata – analog etwa *Elchtest* – aufgenommen, dann aber ggf. wieder gestrichen werden.

tungen und Zeitschriften würden Wörter wie *Schweinegrippe*, aber auch *Flashmob*, vermutlich nie entdeckt werden, jedenfalls nicht in hoher Frequenz. Innovative korpusanalytische Verfahren ermöglichen es, solche Wörter ‘zeitnah’ zu ermitteln und ihre Entwicklung zu verfolgen, etwa dadurch, dass Entstehungsprozesse von Neologismen mit bestimmten Suchalgorithmen gezielt aufgespürt werden (z.B. in den Phrasen *unter ... versteht man*, *dies bedeutet ...*). Bei diesen Neologismen wird deutlich sichtbar, wie Worthäufigkeit und das Vorkommen eines Wortes im Korpus gesellschaftliche oder auch technische Entwicklungen widerspiegeln.

3.2 *Gigaliner* – die Entstehung eines Neologismus

Mit Hilfe eines Korpus kann auch verfolgt werden, wie ein Neologismus entsteht, zum Beispiel der *Gigaliner*. Im Jahr 2006, als die Idee des Mammutfahrzeugs in Deutschland aufkam, stehen verschiedene Begriffe in Konkurrenz zueinander: *Riesen-Lkw*, *Eurocombi*, *Monstertruck*, *XXL-Brummi* und *Gigaliner*. Im Korpus zeigen Metasignale an, dass das Wort noch nicht fest etabliert ist: der Zusatz *sogenannt* (z.B. *sog. Gigaliner*), Anführungszeichen, sonstige Auszeichnungen. Das Ausbleiben von Metasignalen – plus eine stabile Frequenz – ist dann ein Anzeichen dafür, dass sich der Neologismus etabliert hat.

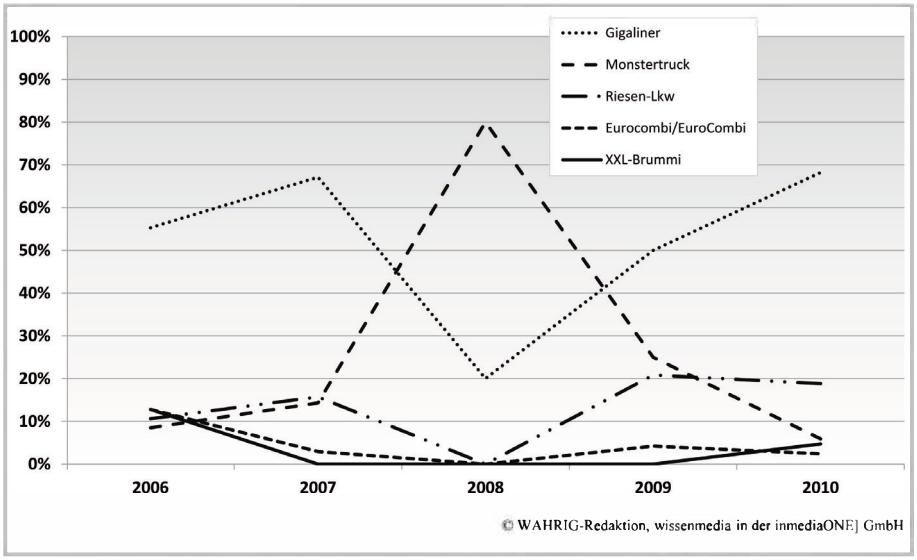


Abb. 4: Die Entstehung eines Neologismus am Beispiel von *Gigaliner*

3.3 *Rettungsschirm*: Semantische Modifikation und Entstehen einer neuen Bedeutung

Es wurde gezeigt, wie ein Neologismus entsteht und wie man ihn anhand eines geeigneten Korpus erkennen und seine Entwicklung verfolgen kann. Mit Hilfe vor allem der Anwendungsbeispiele im Korpus sind auch Bedeutungsveränderungen und -erweiterungen aufspürbar. Dies belegt das Wort *Rettungsschirm*. Bis 2007 taucht das Wort lediglich in der Bedeutung ‘Fallschirm’ auf. Bereits 2008 bis 2009 wird der Begriff häufiger im Sinne von ‘finanzielle Hilfe’ gebraucht. Mit den verschiedenen ‘Rettungspaketen’ für verschuldete EU-Staaten ist dann fast ausschließlich die figurative Bedeutung vorherrschend.

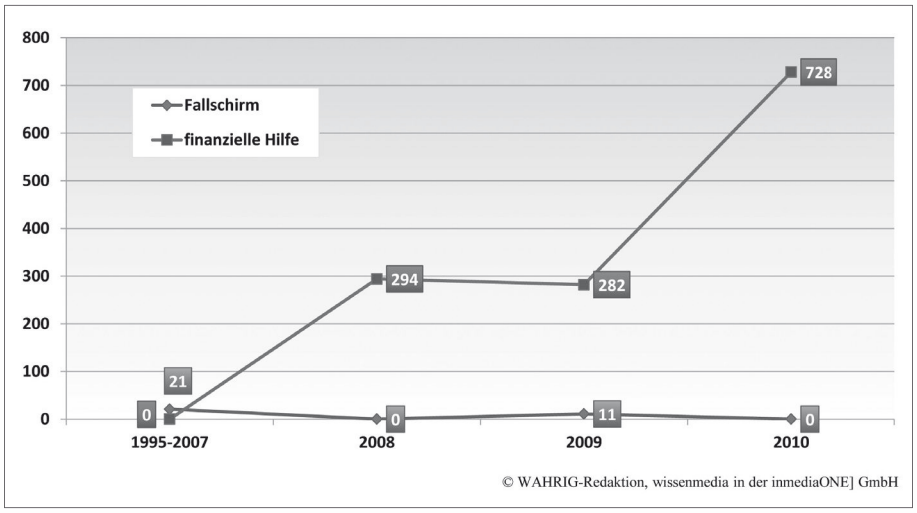


Abb. 5: Semantische Modifikation und Entstehen einer neuen Bedeutung: *Rettungsschirm*

Abschließend soll auf einen der produktivsten Bereiche für Neologismen eingegangen werden, aber auch einen der schnelllebigsten. Dies verdeutlicht eine Korpusanalyse mit Momentaufnahmen aus 70 Jahren.

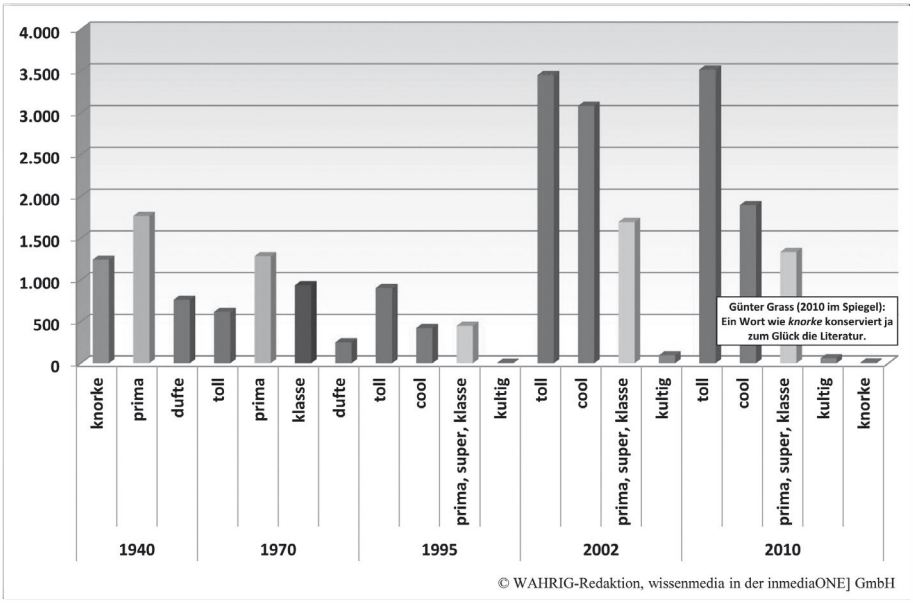


Abb. 6: Neologismen und Jugendsprache

3.4 Neologismen in der Jugendsprache

Im Jahr 2010 zeigt das Korpus elf Begriffe für *sehr gut: knorke*, *dufte*, *prima*, *klasse*, *toll*, *super*, *cool*, *geil*, *megageil*, *endgeil*, *kultig*. Um ihren 'Werdegang' zu verfolgen, wurde ein begrenztes literarisches Korpus des 20. Jahrhunderts ausgewertet.[12] Auf der Jahresmarke 1940 kristallisieren sich drei frequente jugendsprachliche Begriffe für *sehr gut* heraus: *knorke*, *prima* und *dufte*, *prima* als das frequenteste. 30 Jahre später, im Jahr 1970, sind zwei der Begriffe von 1930 immer noch nachzuweisen: *prima* und *dufte*, allerdings mit verminderter Frequenz, *knorke* ist praktisch verschwunden. Hinzugekommen sind *toll* und *klasse*, die beide ihren Platz über die nächsten 40 Jahre behaupten. Im Jahr 1995 mit dem Anfang des systematischen digitalen Korpusaufbaus ist *dufte* verschwunden bzw. nur noch irrelevant in Bezug auf Berliner Varietäten nachzuweisen, hinzugekommen sind *cool*, *super* und *kultig*. Auffällig beim Sprung zum Jahr 2010 ist, dass *cool* sich längst als stabiler Neologismus der Jugendsprache etabliert hat, das zeigt die hohe Zahl der Belege (31.189). Aber *toll* läuft ihm immer noch den Rang ab. Doch auch die 'relativ alten' Begriffe

[12] Dieses Korpus enthält z.B. Bücher und andere literarische/publizistische Dokumente der 1930er und 1940er Jahre (Zeitschriften zum Bereich Kabarett und zu Schriftstellern und Zeichnern wie Döblin oder Zille). Aufgrund juristischer Auflagen darf es aber nicht vollständig genutzt und ausgewertet werden.

prima, *super*, *klasse* haben eine gleichbleibend hohe Frequenz. Will man also wirkliche Neologismen eines Zeitraums ermitteln, kann nur eine erweiterte Suche helfen: Kookkurrenzanalysen unter Einbeziehung von Umgebungswörtern, hier eines anderen – vor allem in der gesprochenen Sprache – nachzuweisenden Neologismus, nämlich: *voll* als graduierendes, adverbial gebrauchtes Adjektiv, wie in *voll kultig*, *voll cool*, *voll geil*, neben *voll krass* und *voll ätzend.* Hier zeigen sich noch erhebliche Desiderate in der gegenwärtigen Korpusarbeit, auch werden die Grenzen eines Zeitungskorpus aus Texten ‘professioneller Schreiber’ deutlich – jugendsprachliche Wortschatzentwicklung ist nur ansatzweise ablesbar, da nicht verschriftlicht.[13]

4. Schreibbeobachtung im Rat für deutsche Rechtschreibung

Stand bis hierher vorwiegend die Sprachentwicklung im Vordergrund der Betrachtung, so soll es nun um die Beobachtung der Schreibentwicklung im Rat für deutsche Rechtschreibung gehen. Diese Aufgabe nimmt mit Hilfe korpusanalytischer Methoden und Instrumentarien die *AG Korpus* wahr – auf der Basis der Korpora von *DUDEN*, IDS und *WAHRIG*, den drei größten Korpora zur deutschen Gegenwartssprache. Für die 2. Amtsperiode des Rates bis 2016 hat die AG ein Konzept entwickelt, wie Schreibbeobachtung langfristig und systematisch betrieben werden kann. Im Mittelpunkt der Beobachtung und Auswertung steht auch hier die Zielgruppe ‘professionelle Schreiber’. Mit Bezug auf die bereits 2006 entwickelte Methodik soll sich die Analyse im Wesentlichen an zwei Fragestellungen orientieren:

1) Akzeptanz der geltenden Rechtschreibregelung und
2) Präferenz bei Variantenschreibungen.

Kernkriterien für eine systematische Analyse sind auch hier wie für die Wörterbucharbeit:

1) Repräsentativität und relative Vollständigkeit der Untersuchungsgegenstände: Die Analyse sollte alle relevanten Bereiche und Komplexe des amtlichen Regelwerks und alle wichtigen Fälle des amtlichen Wörterverzeichnisses abdecken.
2) Repräsentatives Vorkommen im deutschen Wortschatz: Zumindest mittlere Frequenz muss gegeben sein, da der ‘Usus’ analysiert werden soll.

[13] Zur Problematik der Analyse von Soziolekten, vor allem der Jugendsprache, vgl. Elsen (2002, bes. S. 138).

3) Die Ausgewogenheit der Auswahl von allgemeinsprachlichem Wortschatz und einschlägigem Fachwortschatz sollte gewährleistet sein.

4) Die Realisierbarkeit im Rahmen korpusanalytischer Instrumentarien ist im Auge zu behalten.

Als perspektivisches Ziel der Untersuchungen wird angestrebt:

- die Anpassung einer Regel an den beobachteten Gebrauch mit Präzisierung oder Neufassung einer Regel,
- die Neuzuordnung einzelner Lexeme von einer vorhandenen Regel zu einer anderen,
- die Streichung von Einzelschreibungen, die nicht dem Schreibgebrauch entsprechen, bzw. die Neuzulassung von Varianten,
- ggf. die Vereinheitlichung von Schreibungen in den Wörterbüchern im Sinne der Einheitlichkeit der Rechtschreibung im deutschen Sprachraum,
- die Entwicklung von Paradigmen und Pilotuntersuchungen zu orthografischen Fehlerschwerpunkten für den Bereich Schule.

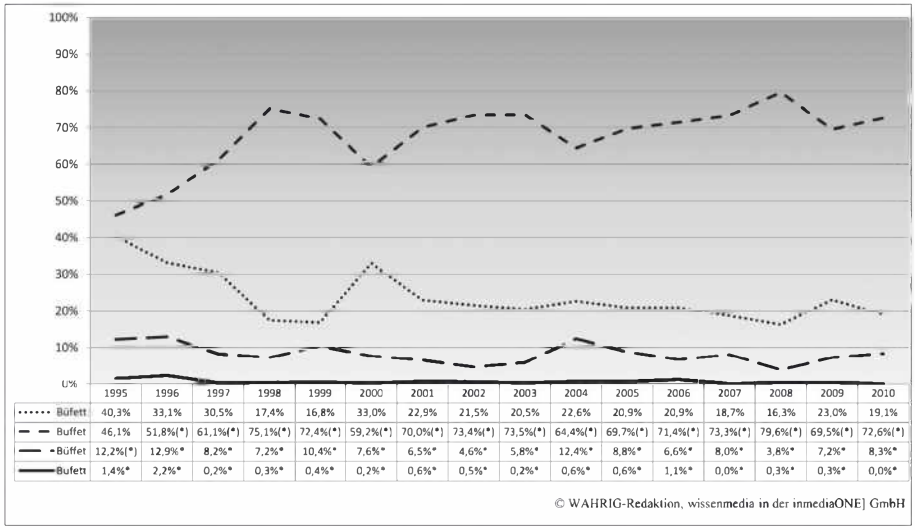
4.1 Die Auswahl der Untersuchungsgegenstände

Hier spielt eine große Rolle, ob die Rechtschreibphänomene für eine Korpusanalyse auf gegenwärtigem Stand systematisch untersucht werden können. Dies ist von den sechs Bereichen des amtlichen Regelwerks bei der Laut-Buchstaben-Zuordnung (LBZ) einschließlich Fremdwörtern, der Getrennt- und Zusammenschreibung (GZS) einschließlich der Schreibung mit Bindestrich und der Groß- und Kleinschreibung (GKS) der Fall. Die LBZ ist in allen Bereichen am besten ‘flächendeckend’ zu analysieren. Zum einen ist die Regelung in allen Bereichen der LBZ seit 1996 konstant geblieben, was zuverlässige Ergebnisse verspricht. Vor allem aber weist die LBZ lediglich orthografische Variation, keine semantisch begründete auf. Als Beispiel dafür, wie offizielle Regeln und tatsächlicher Schreibgebrauch auseinanderdriften können, kann eine Analyse zu den Kategorien Regelakzeptanz und Präferenz von Varianten dienen.

4.2 Varianten bei Fremdwörtern: Diskrepanz zwischen Norm und Usus

Nach amtlichem Regelwerk von 1996 war die fremdsprachige Schreibung *Buffet* nur noch in Österreich und der Schweiz zugelassen. Die Wörterbücher führten sie z.T. trotzdem auf, weil sie schon damals gängiger war als die inte-

grierte Form. Ab 1995 steigend bis 2010 weist sie, nicht nur in den österreichischen Quelltexten, durchgängig die höheren Belegraten auf (obere gestrichelte Linie). Die seit 1996 allein zugelassene integrierte Schreibung *Büfett* (gepunktete Linie) dagegen bewegt sich langfristig fast auf dem gleichen niedrigen Niveau der seit 1996 auch für Österreich nicht mehr zugelassenen Variante *Büffet*. Dies zeigt, dass bei Schreibweisen, die der Logik, z.B. der gängigen Aussprache im Deutschen, widersprechen, die Nachvollziehbarkeit und damit die Akzeptanz von integrierten Schreibungen bei den Schreibenden offenbar gering ist. Konsequenz wäre, die fremdsprachige Schreibung im gesamten deutschen Sprachraum zuzulassen.[14]



	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
Büfett	40,3%	33,1%	30,5%	17,4%	16,8%	33,0%	22,9%	21,5%	20,5%	22,6%	20,9%	20,9%	18,7%	16,3%	23,0%	19,1%
Buffet	46,1%	51,8%(*)	61,1%(*)	75,1%(*)	72,4%(*)	59,2%(*)	70,0%(*)	73,4%(*)	73,5%(*)	64,4%(*)	69,7%(*)	71,4%(*)	73,3%(*)	79,6%(*)	69,5%(*)	72,6%(*)
Büffet	12,2%(*)	12,9%*	8,2%*	7,2%*	10,4%*	7,6%*	6,5%*	4,6%*	5,8%*	12,4%*	8,8%*	6,6%*	8,0%*	3,8%*	7,2%*	8,3%*
Bufett	1,4%*	2,2%*	0,2%*	0,3%*	0,4%*	0,2%*	0,6%*	0,5%*	0,2%*	0,6%*	0,6%*	1,1%*	0,0%*	0,3%*	0,3%*	0,0%*

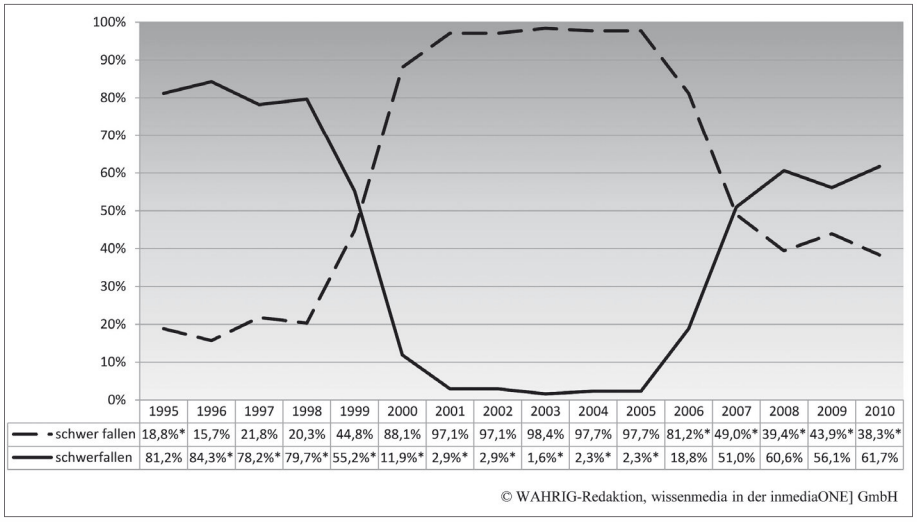
Abb. 7: Variantenschreibungen – Diskrepanz zwischen Norm und Usus: *Büfett* vs. *Buffet*

4.3 Akzeptanz der geltenden Regelung: Getrennt- und Zusammenschreibung

Schwieriger stellt sich die Situation bei der GZS dar. Hier konnten in der 1. Amtsperiode des Rats nur eingeschränkt Analysen gemacht werden, da mit der Neufassung von 2006 zahlreiche Schreibungen neu geregelt wurden.

[14] Ein ähnlicher Fall liegt bei den integrierten Variantenschreibungen *Kreme* oder *Butike* vor, die 2010 vom Rechtschreibrat gestrichen wurden. Das entgegengesetzte Ergebnis, also dass integrierte Schreibungen mehrheitlich gut angenommen werden, zeigt sich bei Lemmata mit den Bestandteilen *phon*, *phot* und *graph*, bei denen die Variante mit *f* seit 1996 regelhaft zugelassen ist. Das Nebeneinander der beiden Formen hier entspricht offenbar dem normalen Prozess der Fremdwortintegration. Dazu ausführlich Krome (2011).

Größtes Problem ist, dass semantische Aspekte häufig eine wichtige Rolle spielen, dieser Punkt ist in der Neuregelung noch gestärkt worden. Dazu sollen die Beispiele *schwerfallen* und *leichtfallen* – Verbindungen von Adjektiv und Verb – untersucht werden, beide in übertragener Bedeutung.

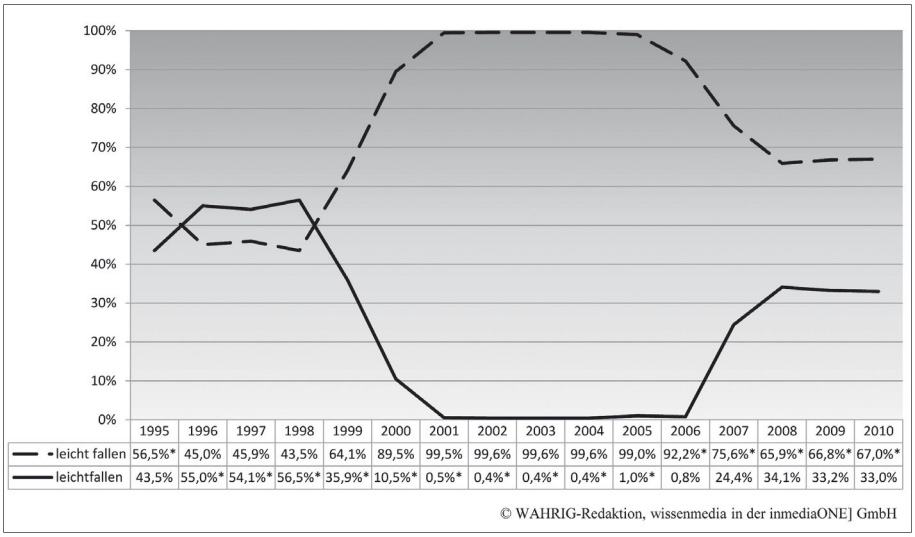


	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
schwer fallen	18,8%*	15,7%	21,8%	20,3%	44,8%	88,1%	97,1%	97,1%	98,4%	97,7%	97,7%	81,2%*	49,0%*	39,4%*	43,9%*	38,3%*
schwerfallen	81,2%	84,3%*	78,2%*	79,7%*	55,2%*	11,9%*	2,9%*	2,9%*	1,6%*	2,3%*	2,3%*	18,8%	51,0%	60,6%	56,1%	61,7%

Abb. 8: Schreibbeobachtung zur Getrennt- und Zusammenschreibung: *schwerfallen* vs. *schwer fallen*

Bei *schwerfallen* ist das Ergebnis eindeutig: die Wiedereinführung der Zusammenschreibung greift. Auch vor der Reform wurde zu rund 80% richtig zusammengeschrieben. Ab 1998 zeigt die Reform dann Wirkung. Die Gebräuchlichkeit der neu eingeführten Getrenntschreibung verdrängt die alte Zusammenschreibung fast vollständig, aber in der Öffentlichkeit gibt es starke Kritik an der bedeutungsfernen Getrenntschreibung. Seit 2006 ist ein deutlicher Anstieg der Zusammenschreibung nach ihrer Wiederzulassung festzustellen, 2009 fast wieder auf Vorreformniveau. Dieser Fall wird weiter beobachtet werden, z.B. um das neue Regelkriterium der Idiomatisierung zu prüfen.

Bei *leichtfallen* ist das Ergebnis nicht analog, hier zeigt sich auch 2010 noch eine deutliche Bevorzugung der Getrenntschreibung. Dies könnte daran liegen, dass das Verb auch in konkreter Bedeutung gebraucht werden kann, woraufhin alle Belege eigens geprüft werden müssen. Aber schon die bloße Tatsache, dass ein Wort konkret oder übertragen gebraucht werden kann, kann die ‘Intuition der Schreibenden’ offenbar beeinflussen. Hier wären sprachtechnologisch bessere Möglichkeiten zu kontextsensitiven – syntaktischen und semantischen – Analysen sehr von Nutzen. Dies trifft auch auf einige Komplexe der Groß- und Kleinschreibung zu.



	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
leicht fallen	56,5%*	45,0%	45,9%	43,5%	64,1%	89,5%	99,5%	99,6%	99,6%	99,6%	99,0%	92,2%*	75,6%*	65,9%*	66,8%*	67,0%*
leichtfallen	43,5%	55,0%*	54,1%*	56,5%*	35,9%*	10,5%*	0,5%*	0,4%*	0,4%*	0,4%*	1,0%*	0,8%	24,4%	34,1%	33,2%	33,0%

Abb. 9: Schreibbeobachtung zur Getrennt- und Zusammenschreibung: *leichtfallen* vs. *leicht fallen*

4.4 Akzeptanz der geltenden Regelung: Groß- und Kleinschreibung

Die Korpusanalyse zeigt, dass in der GKS die formalen Kriterien der Rechtschreibreform 1996 in vielen Fällen offenbar greifen. So wird die Neuregelung z.B. bei Fügungen aus Präposition, Artikel und substantiviertem Partizip nahezu zu 100% akzeptiert und besser angenommen als die Vorreformregelung nach Bedeutungskriterien (wie etwa bei dem Fallbeispiel *auf dem Laufenden sein*). Schwieriger liegt der Fall bei den sog. Nominationsstereotypen, z.B. bei *Schwarzes Brett* oder *Rote Karte*. Hier muss der Schreibende in jedem Einzelfall entscheiden, ob eine konkrete oder eine übertragene Bedeutung vorliegt oder ob der Begriff zur Fachsprache gehört. Es ist daher mit den gegenwärtig zur Verfügung stehenden Mitteln nur schwer auszumachen, ob die geltende Regelung der Bedeutungsdifferenzierung angenommen wird oder nicht.

5. Die Chancen 'zeitungskorpusbasierter' Lexikographie

Sowohl die Schreib- wie auch die Wortschatzbeobachtung vor allem in der Neologismenarbeit zeigt, welch große Chancen das Korpus, ein Zeitungskorpus, für die moderne Lexikographie bietet. Denn hier werden Sprache und Rechtschreibung ständig reflektiert. Neue Ausdrücke müssen eingeführt werden, da sie beim Leser nicht immer als bekannt vorausgesetzt werden können. So können bei der Sprachbeobachtung im Rahmen von Wortschatzanalysen fehlende

Wörter ermittelt und zeitnah neue Wörter aufgespürt werden. Veraltete Wörter können im Gegenzug so gekennzeichnet oder gestrichen werden. Im Jahrgangsvergleich können Neologismen im Hinblick auf Vorkommen, Bedeutung und Schreibung beobachtet werden, auch im Umfeld von weiteren Neologismen. Durch Ermittlung von Kollokatoren kann ferner der Bedeutungswandel von Wörtern und Wendungen verfolgt werden. Und im Zuge der Schreibbeobachtung im Rat für Rechtschreibung schließlich kann eine Prüfung von Schreibweisen und dadurch von orthografischen und grammatischen Regularitäten vorgenommen werden.

All dies ermöglicht differenzierte Aufschlüsse zu authentischem Sprachgebrauch. Die Methoden der reinen Frequenzanalyse sind inzwischen deutlich verbessert und modifiziert worden. Es bleiben aber klare Defizite im Bereich der Kollokations- und Kontextanalyse. Die gezeigten Beispiele haben deutlich gemacht, dass viele Fragen, bei denen semantische Bezüge eine Rolle spielen, mit Hilfe derzeit zur Verfügung stehender Methoden nicht oder nur unter großem Aufwand geklärt werden können. Wie also könnte das Korpus mittel- und langfristig weiterentwickelt werden?

6. Korpusanalyse – Desiderate und Perspektiven

1) Textliche Grundlagen

   a) Im Hinblick auf Fehleranalysen sind die bisher zugrundeliegenden Korpora defizitär: Ein Korpus, das vorwiegend aus Texten ‘professioneller Schreiber’ besteht, spiegelt nur begrenzt Fehlschreibungen, da es zum großen Teil auf Korrekturprogrammen aufsetzt. Eine Lösungsmöglichkeit wäre der Aufbau eines Subkorpus aus Internettexten wie Blogs und E-Mails.

   b) Bei Wortschatzanalysen kann etablierter oder veralteter Wortschatz durch Frequenzbeobachtungen häufig nicht als wichtig, veraltet, überarbeitungsbedürftig erkannt werden. Hier wäre es möglich, eine Auswahl an historischen Texten einzuspeisen: Zeitungstexte, aber auch Literatur.

   c) Im Hinblick auf Stilanalysen ist die Umgangssprache unterrepräsentiert. Jugendwortschatz etwa kann nur in Ansätzen reflektiert werden. Denkbar wäre der Aufbau eines Korpus gesprochener Sprache, ggf. auch auf der Basis anderer Medien wie Radio und Fernsehen. Internettexte könnten ebenfalls eine gute Grundlage sein, ebenso wie ein Teilkorpus repräsentativer Jugendliteratur.

2) Ausgereifte Methoden der Kontext- und Kollokationsanalyse

In diesem Bereich besteht der größte Optimierungsbedarf: Die sprachtechnologischen Instrumentarien sind unzulänglich im Hinblick auf eine syntaktische und semantische Auswertung, etwa hinsichtlich der Wortartenbestimmung. So ist auch die Zeichensetzung bisher nicht korpusanalytisch zu untersuchen. Was ist hier zu tun?

Die grundlegenden Desiderate, die beim jetzigen Stand empirischer korpusanalytischer Forschung noch vorhanden sind, sowie der hohe finanzielle, technische und personelle Aufwand, den eine grundlegende Optimierung der Methoden und Instrumentarien erforderlich macht, andererseits aber auch die hoffnungsvollen Ansätze und Ergebnisse der Kooperation verschiedener `Korpuspartner` im Rat für deutsche Rechtschreibung legen eine übergreifende Zusammenarbeit nahe. Ziel einer solchen Kooperation von Partnern, die ähnlich gelagerte Korpora aufgebaut haben, wäre eine umfassende Dokumentation des Schreibgebrauchs mit Hilfe innovativer computerlinguistischer und sprachtechnologischer Werkzeuge und Methoden für die Korpusbearbeitung und für vergleichende Analysen in einer neuen, übergreifenden Forschungsinfrastruktur. Zu erhoffen wären entscheidende Synergie-Effekte, um viele der oben angesprochenen Probleme zu lösen, so dass die einzelnen Korpora noch effektiver und gewinnbringender ausgewertet werden könnten. Darüber hinaus könnte eine Zusammenarbeit mit Institutionen, die andersgeartete Korpora entwickelt haben, sinnvoll sein, so dass wertvolles Auswertungsmaterial zumindest teilweise auch weiteren Nutzergruppen zur Verfügung gestellt werden könnte[15] – für wissenschaftliche Folgeprojekte beispielsweise zur Fremdwort- oder Neologismenforschung und damit zu einer noch präziseren und umfassenderen Beobachtung des Sprach- und Schreibgebrauchs.

Literatur

Quellen/Korpustexte

Brockhaus WAHRIG Deutsches Wörterbuch. Hrsg. von Renate Wahrig-Burfeind. Gütersloh/München 2011.

Brockhaus WAHRIG Die deutsche Rechtschreibung. Hrsg. von der WAHRIG-Redaktion. Gütersloh/München 2011.

[15] Dieses Desiderat zeigen die meisten vorliegenden korpusanalytischen Studien: Auf ein zielgerichtet aufgebautes Gesamtkorpus kann häufig nicht zurückgegriffen werden. Auch besteht noch zu wenig Verbindung zwischen universitär-wissenschaftlichem und empirisch arbeitendem Bereich.

COSMAS II: http://www.ids-mannheim.de/cosmas2 (Stand: 16.02.2012).

Lemnitzer, Lothar: Die Wortwarte. Wörter von heute und morgen. Eine Sammlung von Neologismen: http://www.wortwarte.de (Stand: 16.02.2012).

WAHRIG Textkorpus^{digital}: http://ww.brockhaus.de/wahrig (Stand: 18.02.2012).

Wortschatz Universität Leipzig: http://wortschatz.uni-leipzig.de (Stand: 16.02.2012).

Wissenschaftliche Literatur

Eichinger, Ludwig M. (2010): Der durchschnittliche Linguist und die Daten. Eine Annäherung. In: Kratochvílová/Wolf (Hg.), S. 27-51.

Elsen, Hilke (2011): Neologismen. Formen und Funktionen neuer Wörter in verschiedenen Varietäten des Deutschen. 2. Aufl. (= Tübinger Beiträge zur Linguistik 477). Tübingen.

Elsen, Hilke (2002): Neologismen in der Jugendsprache. In: Muttersprache 112, 2, S. 136-154.

Elsen, Hilke/Dzikowicz, Edyta (2005): Neologismen in der Zeitungssprache. In: Deutsch als Fremdsprache 42, 2, S. 80-85.

Herberg, Dieter/Kinne, Michael/Steffens, Doris (2004): Neuer Wortschatz. Neologismen der 90er Jahre im Deutschen. Berlin/New York.

Kallmeyer, Werner/Zifonun, Gisela (Hg.) (2007): Sprachkorpora – Datenmengen und Erkenntnisfortschritt. Jahrbuch 2006 des Instituts für Deutsche Sprache. Berlin/ New York.

Klosa, Annette (2007): Korpusgestützte Lexikographie: besser, schneller, umfangreicher? In: Kallmeyer/Zifonun (Hg.), S. 105-122.

Klosa, Annette (2010): Chancen und Probleme korpusgestützter Lexikografie. Am Beispiel deutschsprachiger Online-Wörterbücher. In: Kratochvílová/Wolf (Hg.), S. 103-115.

Kratochvílová, Iva/Wolf, Norbert Richard (Hg.) (2010): Kompendium Korpuslinguistik. Eine Bestandsaufnahme aus deutsch-tschechischer Perspektive. Heidelberg.

Krome, Sabine (2010): Die deutsche Gegenwartssprache im Fokus korpusbasierter Lexikographie. Korpora als Grundlage moderner allgemeinsprachlicher Wörterbücher am Beispiel des WAHRIG Textkorpus^{digital}. In: Kratochvílová/Wolf (Hg.), S. 117-134.

Krome, Sabine (2011): Variantenschreibungen bei Fremdwörtern: Darstellung und Begründung. Empirische Schreibbeobachtung auf der Grundlage korpusbasierter Lexikographie. In: Mitteilungen des Deutschen Germanistenverbandes. Rechtschreibung 58, 1, 2011, S. 36-50.

Lemnitzer, Lothar/Zinsmeister, Heike (2010): Korpuslinguistik. Eine Einführung. 2. Aufl. Tübingen.

Lüdeling, Anke (2007): Das Zusammenspiel von qualitativen und quantitativen Methoden in der Korpuslinguistik. In: Kallmeyer/Zifonun (Hg.), S. 28-48.

Meger, Andreas W. (2010): Makro- und mediostrukturelle Aspekte in Neologismenwörterbüchern. Ein Beitrag zur Theorie und Praxis der Neografie des Polnischen, Russischen, Tschechischen und Deutschen. Univ. Diss. Mannheim.

Mindt, Ilka (2010): Methoden der Korpuslinguistik: Der korpus-basierte und der korpusgeleitete Ansatz. In: Kratochvílová/Wolf (Hg.), S. 53-65.

Scherer, Carmen (2006): Korpuslinguistik. (= Kurze Einführungen in die germanistische Linguistik 2). Heidelberg.

Quasthoff, Uwe (2007): Deutsches Neologismenwörterbuch. Neue Wörter und Wortbedeutungen in der Gegenwartssprache. Berlin/New York.

URL 1: Bubenhofer, Noah: Einführung in die Korpuslinguistik: Praktische Grundlagen und Werkzeuge (2006-2012). http://www.bubenhofer.com/korpuslinguistik (Stand: Februar 2012).

Wolf, Norbert Richard (2010): Korpora in der Korpuslinguistik. In: Kratochvílová/Wolf (Hg.), S. 17-25.

Wolf-Bleiß, Birgit (2009): Neologismen – Sprachwandel im Bereich der Lexik. In: Siehr, Karl-Heinz/Berner, Elisabeth (Hg.): Sprachwandel und Entwicklungstendenzen als Themen im Deutschunterricht: fachliche Grundlagen – Unterrichtsanregungen – Unterrichtsmaterialien. Potsdam, S. 83-101.