Wolfgang Klein / Alexander Geyken

Das Digitale Wörterbuch der Deutschen Sprache (DWDS)

Abstract
No area in the study of human languages has a longer history and a higher practical significance than lexicography. The advent of the computer has dramaticually changed this discipline in ways which go far beyond the digitisation of materials in combination with efficient search tools, or the transfer of an existing dictionary onto the computer. They allow the stepwise elaboration of what is called here Digital Lexical Systems, i.e., computerized systems in which the underlying data - in form of an extendable corpus - and description of lexical properties on various levels can be efficiently combined. This paper discusses the range of these possibilities and describes the present form of the German „Digital Lexical System of the Academy", a project of the Berlin-Brandenburg Academy of Sciences (www.dwds.de).

Das Grimmsche Wörterbuch ist unzweifelhaft ein großes nationales Werk deutschen Gelehrtenfleißes, und es würde aufs tiefste zu beklagen sein, wenn dasselbe unvollendet bleiben müßte.
Otto von Bismarck (1868)[1]

1 Einleitung

Es hat die Vollendung erlebt, und auch wieder nicht, denn als im Jahre 1960 die letzte Lieferung des *Deutschen Wörterbuchs* erschien, da war längst deutlich, dass weite Teile dieses gewaltigen Werks völlig überholt waren: Es hat sich, je weiter es voranschritt, zugleich von

[1] In einer Vorlage für den Bundesrat, zitiert nach Stackmann (2002), S. 251.

DOI 10.1515/lexi.2010.006

der Erfüllung seiner Aufgabe, den deutschen Wortschatz in seiner geschichtlichen Entwicklung umfassend darzustellen, immer weiter entfernt. Dies gilt besonders für die noch von Jacob und Wilhelm Grimm bearbeiteten Teile mit den Buchstaben A – F, die bei Abschluss des Werkes schon ein Jahrhundert alt waren. So wurde denn zwischen der Deutschen Akademie der Wissenschaften in Berlin (jetzt BBAW) und der Göttinger Akademie der Wissenschaften vereinbart, zunächst einmal diese Teile gemeinsam neu zu bearbeiten; was mit den Buchstaben G – Z geschehen sollte, würde man sehen. Die Beratungen, die zu diesem Beschluss geführt haben, waren durchaus kontrovers, denn zum einen war klar, dass auch andere Teile längst überholt waren, und zum anderen war die Befürchtung nicht ganz von der Hand zu weisen, dass auch die Neubearbeitung eben dieser Buchstaben über die ursprünglich 15 avisierten Jahre hinaus dauern könnte. Das hat sich bestätigt. Die erste Lieferung ist 1965 erschienen, mit dem Abschluss ist nun, einigermaßen sicher, im Jahre 2013 zu rechnen. Was soll danach geschehen?

Man könnte so weitermachen wie bisher. Dagegen sprechen zwei Gründe, ein negativer und ein positiver:

A. Wenn die Neubearbeitung der Buchstaben A – F rund fünzig Jahre gedauert hat, dann wird, wenn man so fortfährt, der Abschluss überschlägig noch weitere 200 Jahre auf sich warten lassen. Niemand wird dies finanzieren wollen, und viel schlimmer: Je weiter die Arbeiten vorangehen, umso überalteter wird das Werk selbst in den bereits erschienenen Teilen. Wir wissen nicht, was die nächsten zweihundert Jahre uns bringen, aber man darf sicher annehmen, dass sich die deutsche Sprache und insbesondere der deutsche Wortschatz weiter entwickeln.

B. Die technische Entwicklung hat uns eine Reihe von Möglichkeiten eröffnet, den Wortschatz einer Sprache anders als in Form eines konventionellen Wörterbuchs darzustellen: An die Stelle eines gedruckten Wörterbuchs tritt ein über Internet zugängliches *Digitales Lexikalisches System*, das Belege und lexikalische Analyse miteinander verbindet, in dem Informationen leicht und schnell nachschlagbar sind und das sich jederzeit flexibel erweitern und korrigieren lässt. Es leistet alles, was ein klassisches Wörterbuch leistet, hat darüber hinaus aber einen erheblichen Mehrwert; beispielsweise kann man sich die Aussprache vorsprechen lassen, statt sie über eine Lautschrift erschließen zu müssen. Dies macht gedruckte Wörterbücher für spezifische Zwecke nicht überflüssig. Sie lassen sich aber jederzeit und ohne größeren Aufwand aus dem eigentlichen auf Computer verfügbaren lexikalischen System ableiten.

Im Jahr 2005 hat eine kleine Arbeitsgruppe[2] an der Berlin-Brandenburgischen Akademie der Wissenschaften den Plan für ein solches Wörterbuchsystem entwickelt. Er fußt auf einem etwas älteren, gleichfalls DWDS genannten Vorhaben, in dem es jedoch nur um ein Wörterbuch der Gegenwartssprache ging.[3] Das neue Vorhaben, das seit 2007 im Langzeit-

---

[2] Ihr gehören an: Manfred Bierwisch, Alexander Geyken, Wolfgang Klein (Leitung), Wolf-Hagen Krauth, Hartmut Schmidt, Angelika Storrer.

[3] Mit den Vorarbeiten zu diesem *Digitalen Wörterbuch der deutschen Sprache des 20. Jahrhunderts (DWDS)*, das auf eine Anregung von Hartmut Schmidt zurückgeht, wurde bereits 1999 begonnen. Der Schwerpunkt der Arbeiten lag zunächst auf dem Aufbau eines Korpus (vgl. Abschnitt 4). Als

programm der Union der Akademien gefördert wird, wird hingegen auch die historische Wortforschung, und damit auch den *Grimm*, integrieren. Im Folgenden werden Hintergrund, Ziele und derzeitiger Stand des Vorhabens beschrieben; über Internet ist es unter http://www.dwds.de bzw. unter http://beta.dwds.de/ zugänglich.[4]

2 Probleme des klassischen Wörterbuchs

Das Ziel der Lexikographie ist es, den Wortschatz einer Sprache zu bestimmten Zwecken umfassend oder in bestimmten Ausschnitten zu beschreiben (und, beim zweisprachigen Wörterbuch, ihn zu dem einer anderen Sprache in Bezug zu setzen). Verwirklicht wird dieses Ziel gewöhnlich in Form eines gedruckten Wörterbuchs, in dem die einzelnen lexikalischen Einheiten alphabetisch oder nach einem sonstigen Prinzip geordnet beschrieben werden. Das ist uns seit Calepinus' Tagen so selbstverständlich, dass man sich nur schwer ein anderes Vorgehen denken kann. Aber dieser Weg, den Wortschatz einer Sprache darzustellen, hat unweigerlich mit einer ganzen Reihe von Problemen zu kämpfen, die durch das Format selbst diktiert sind. Die wichtigsten davon sind:

2.1 Probleme des Umfangs

Der Raum, der in einem Wörterbuch zur Verfügung steht, ist beschränkt. Dies wirkt sich auf die Auswahl der lexikalischen Einheiten wie auf die Beschreibung der verschiedenen Eigenschaften einer Einheit selbst stark aus:

(a) Auch das umfangreichste gedruckte Wörterbuch kann nur einen kleinen Teil der tatsächlichen lexikalischen Einheiten aufnehmen. Das [1]DWB umfaßt rund 330 000 Einträge. Im DWDS-Korpus, der derzeit repräsentativsten Zusammenstellung deutschsprachiger Texte des 20. Jahrhunderts (s.u. Abschnitt 4 *B*), finden sich jedoch allein für diesen Zeitraum etwa 4 Millionen verschiedene Wörter (in etwa 6 Millionen Flexionsformen).[5] Viele davon sind Komposita wie *Bahnreise* oder *Parklücke,* die nicht jeder als selbständige lexikalische Einheiten rechnen würde. Allerdings ist es nicht so, als könnte man die Bedeutung von *Parklücke* ohne weiteres aus der Bedeutung von *Park* und *Lücke* herleiten. Auch sind alle syntaktisch komplexen, aber ihrer Bedeutung nach lexikalisierten oder halblexikalisierten Ausdrücke, wie etwa *zu Kreuze kriechen, zur Welt bringen*, *ins Zeug legen*, nicht berücksichtigt. Und dies sind nur die lexikalischen Einheiten, die in einem – gemessen an der Gesamtzahl der veröffentlichten Texte – nach wie vor verschwindend kleinen Teil der im

---

sich einige Jahre später die Möglichkeit abzeichnete, ein umfassendes Wörterbuchsystem unter Einschluss der historischen Lexikographie zu entwickeln, wurde der einschränkende Zusatz *des 20. Jahrhunderts* fallengelassen, die Abkürzung DWDS wurde beibehalten.

4 Die Abschnitte 2 und 3 gehen in stark verkürzter Form auf Klein (2004) zurück.

5 Nimmt man Eigennamen, fremdsprachliches Material und Sonderformen hinzu, so kommt man insgesamt auf über 9 Millionen verschiedene Wortformen.

20. Jahrhundert veröffentlichten Texte vorkommen. Kein klassisches Wörterbuch ist in der Lage, diesen Wortschatz auch nur annähernd abzudecken.

(b) Ebenso kann der Platz, der für eine einzelne lexikalische Einheit verfügbar ist, im gedruckten Wörterbuch nicht beliebig erweitert werden. In ihrer langen Geschichte hat die Lexikographie viele Techniken entwickelt, um mit diesem Problem umzugehen – Abkürzungen, Sonderzeichen, Querverweise, besondere Auszeichnungen usw. (siehe Wiegand 1989). Aber diese Möglichkeiten sind begrenzt, und je mehr von ihnen Gebrauch gemacht wird, umso umständlicher wird die Nutzung des Wörterbuchs.

2.2 Mangelnde Flexibilität

Scripta manent: Wenn ein Wörterbuch oder auch nur eine bestimmte Lieferung erst einmal gedruckt ist, dann lässt sich nur noch schwer etwas ändern. Man kann nicht so ohne weiteres neue Einheiten hinzufügen, die erst für die Zeit nach der Drucklegung belegt sind oder die schlichtweg übersehen wurden, und ebenso sind Korrekturen und Ergänzungen des einmal Gedruckten nur in engen Grenzen – etwa durch Nachträge oder Erratazettel – möglich. Der einzige Weg, dieses Problem zu lösen, ist in der Regel eine neue Ausgabe.

2.3 Gebundenheit an den Schriftcode

Die einzelnen Eigenschaften einer lexikalischen Einheit (Aussprache, Bedeutung, morphologische und syntaktische Merkmale usw.) müssen schriftlich dargestellt werden. Davon gibt es einige wenige Ausnahmen, etwa Bebilderungen, denen aber auch aus Gründen des Formats enge Grenzen gesetzt sind. Normalerweise benötigt man einen konventionellen, schriftlichen Code, im einfachsten Fall dieselbe Sprache, deren Wortschatz das Wörterbuch selektiv beschreibt. Für die phonologischen Eigenschaften bedient man sich in der Regel eines besonderen Schriftsystems, von dem man hofft, dass es diese Eigenschaften akkurat wiedergibt, vor allem aber, dass der Benutzer es versteht und die rechte Aussprache daraus ableiten kann. Die Erfahrung lehrt, dass dieses Vorgehen nicht nur sehr fehlerbehaftet ist, sondern allenfalls ein Behelfsmittel, wenn man sich beispielsweise als Sprachlerner die genaue Aussprache aneignen will.

2.4 Zweidimensionalität der Darstellung

Der Wortschatz einer Sprache, gleich in welcher Auswahl er beschrieben wird, ist keine Liste von Wörtern, sondern eine vieldimensionale Struktur, definiert durch die Einheiten und die verschiedenen lexikalischen Relationen zwischen ihnen (vgl. etwa Murphy 2003). Die zweidimensionalen Seiten eines Wörterbuchs können diese Struktur nur sehr unzulänglich abbilden. Das Format des Wörterbuchs erfordert es zunächst einmal, die einzelnen lexikalischen Einheiten linear anzuordnen. Dafür bietet sich – jedenfalls in jenen Sprachen, die eine Alphabetschrift haben – das Alphabet an. Das stößt allerdings an gewisse Beschränkungen, wo man aus Gründen der Platzersparnis einzelne Einheiten zu ‚Nestern' zusam-

menstellt. Gravierender sind die Schwierigkeiten, in die das Alphabetverfahren bei syntaktisch zusammengesetzten lexikalischen Einheiten gerät: Wie ordnet man *den Teufel an die Wand malen* ein? Unter *Teufel*, unter *Wand*, unter *malen,* unter allen dreien? Vor allem aber fasst die alphabetische Reihenfolge nicht zusammen, was der Sache nach zusammengehört. Will man einen anderen Zusammenhang zwischen lexikalischen Einheiten zur Darstellung bringen, so muss man ein neues Wörterbuch – ein Reimwörterbuch, ein Synonymenwörterbuch, ein Valenzwörterbuch – anlegen und in aller Regel, um eine sinnvolle Benutzung zu ermöglichen, durch einen alphabetischen Index ergänzen. Aber auch ein solches Wörterbuch erfasst bestenfalls eine oder zwei der Relationen, die zwischen den lexikalischen Einheiten bestehen, und dies in aller Regel sehr unzulänglich. So werden z.B. einem Bündel phonologischer Eigenschaften ja in aller Regel sehr verschiedene Bündel semantischer Eigenschaften zugewiesen sein. Ein ‚Wort' ist daher immer nur in einer bestimmten Lesart zu einem anderen Wort synonym oder hyponym. Grundsätzlich überwinden kann man diese Beschränkungen nur mit Instrumentarien, die mehrdimensionale Strukturen flexibel abzubilden erlauben: Man benötigt ein System, das verschiedene Wörterbuchtypen integriert.

2.5 Grenzen der Exemplifizierung

Umfassende Wörterbücher fügen aus gutem Grund den einzelnen Einträgen Beispiele für die Verwendung hinzu – entweder selbst erdachte oder Belege aus Texten irgendwelcher Autoren. Diese Exemplifizierung erlaubt, die Aussagen des Lexikographen zu überprüfen, und, viel wichtiger, in vielen Fällen machen erst solche Verwendungsbeispiele die Bedeutungsangaben verständlich; insofern können sie einen wesentlichen Teil der Bedeutungsbeschreibung bilden. Ein Wörterbuch kann dies nur in engen Grenzen leisten; insbesondere kann es nur einen sehr kleinen Teil des Kontextes angeben, in dem die betreffende lexikalische Einheit vorkommt. Viel besser wäre es, die Einträge des Wörterbuchs direkt mit entsprechenden Texten zu verbinden, so dass sich jeder passende Beleg in beliebiger Anzahl und mit beliebig viel Kontext anschauen lassen könnte.

Mit den Mitteln der traditionellen Lexikographie sind die hier skizzierten Probleme und die daraus resultierenden Unzulänglichkeiten nur schwer zu bewältigen. Sie rühren nicht aus mangelnder Kompetenz oder Sorgfalt des einzelnen Bearbeiters, sondern sind systematischer Natur. Umso bewundernswerter muss uns die Leistung erscheinen, die die großen Vertreter des Faches in der Vergangenheit vollbracht haben. Die Entwicklung neuer digitaler Techniken lässt es nun erstmals möglich erscheinen, die klassischen Aufgaben von Lexikologie und Lexikographie in anderer und effizienterer Weise zu lösen. Dies heißt nicht, dass man auf die Kenntnisse und den Sachverstand ausgebildeter Lexikographen verzichten könnte. Im Mittelpunkt der lexikographischen Analyse steht die Beschreibung der Bedeutung, und dazu bedarf es nach wie vor eines Menschen, der die Verwendung eines Wortes in einem bestimmten Kontext zu interpretieren vermag. Aber zum einen lassen sich wesentliche Teile der Arbeit erheblich beschleunigen; dies betrifft vor allem die Erstellung und Nutzung der Belege, auf denen die lexikographische Analyse beruht, ebenso die Einbindung des bisherigen lexikalischen Wissens. Zum andern lässt sich die Gesamtaufgabe sinnvoll in einzelne für sich bearbeitbare Teilaufgaben zerlegen, die eine klar begrenzte Laufzeit haben, getrennt finanzierbar sind und deren jede für sich genommen einen erheblichen Nutzen hat. Zum dritten lässt sich auf diese Weise das vorhandene lexikographische Wissen

leichter in einer durch die Wichtigkeit der Teilaufgaben bestimmten Reihenfolge aufarbeiten und integrieren als bisher. Das Ergebnis ist nicht primär ein gedrucktes Wörterbuch, sondern ein *Digitales Lexikalisches System*, das Belege und lexikalische Analyse miteinander verbindet, in dem Informationen leicht und schnell nachgeschlagen werden können und das sich jederzeit flexibel erweitern und korrigieren lässt. Ein solches Wörterbuchsystem leistet alles, was ein klassisches Wörterbuch leistet, hat darüber hinaus aber einen erheblichen Mehrwert.

3 Digitale Lexikalische Systeme (DLS)

Die meisten bisherigen Computerwörterbücher, ebenso die digitalen Versionen klassischer Wörterbücher wie des OED, des Grimm, des Merriam-Webster, stehen im Prinzip noch in der Tradition des klassischen Wörterbuchs; allerdings bieten sie wesentlich verbesserte Suchmöglichkeiten, und sie lassen sich miteinander verlinken – d.h. man kann von einem ins andere springen. Im Übrigen jedoch unterliegen sie allen Beschränkungen des traditionellen Wörterbuchs. Digitale Lexikalische Systeme sind anders konzipiert. Wie ein Wörterbuch bildet ein solches System die Lexik einer Sprache selektiv ab. Anders ist es vor allem in dreierlei Hinsicht:

A. Es steht nicht auf dem Papier, sondern als elektronische Datenbank im Computer. Allerdings lassen sich jederzeit Papierwörterbücher daraus ableiten.
B. Die lexikalische Analyse ist ständig mit der Datengrundlage – im Wesentlichen ein Textkorpus mit effizienten Suchwerkzeugen – miteinander verknüpft. Ein DLS vereint also die Funktion eines Thesaurus mit der eines klassischen Wörterbuchs.
C. Im Übrigen lässt sich ein DLS durch vier gleichermaßen hässliche Schlagworte kennzeichnen: *Modularität, inkrementelle Funktionalität, kumulative Entwicklung, Methodenpluralität.* Im Grunde sind es alles Abwandlungen ein und derselben Eigenschaft, nämlich einer hohen Flexibilität auf Bearbeiter- und Nutzerseite.

Diese vier unter C genannten Eigenschaften werden im Folgenden näher erläutert.

Modularität

Das gesamte DLS besteht aus einer Reihe von Komponenten, die sich weithin unabhängig voneinander bearbeiten und auch nutzen lassen. Dafür gibt es im Einzelnen unterschiedliche Möglichkeiten; es ist auch nicht ein für alle Mal festgeschrieben. Die einfachste Möglichkeit besteht darin, die einzelnen Module nach den verschiedenen Eigenschaften lexikalischer Einheiten zu ordnen; dann gäbe also es beispielsweise

– ein Modul ‚Aussprache'
– ein Modul ‚Morphologie'
– ein Modul ‚Syntax'
– ein Modul ‚Semantik'
– ein Modul ‚Etymologie'.

Quer dazu liegen Module, die sich nicht an den spezifischen Eigenschaften einer lexikalischen Einheit festmachen lassen. Ein Beispiel sind Angaben über die Verwendungshäufigkeit in bestimmten Texttypen, zu bestimmten Zeiten, durch bestimmte Autoren usw. Die enge Verknüpfung mit dem Belegkorpus lässt dies relativ leicht zu. Man kann daher verfolgen, wie sich ein Wort innerhalb einer Sprachgemeinschaft ausbreitet oder, umgekehrt, wie es allmählich aus dem Sprachgebrauch verschwindet.

In anderer Weise quer dazu liegen Module, die Informationen zu anderen Sprachen hinzufügen, d.h. der traditionell so grundlegende Unterschied zwischen monolingualen, bilingualen und multilingualen Wörterbüchern wird aufgelöst. Der Übergang von einem deutschen zu einem deutsch-französischen Wörterbuch besteht darin, eine neue Komponente an Informationen hinzuzufügen. In ähnlicher Weise kann der innersprachlichen Variation, etwa der Gliederung in Dialekte, Rechnung getragen werden. Es gibt kein pfälzisches Wörterbuch mehr, sondern eine pfälzische Komponente im Gesamtsystem, dessen Einheiten mit denen anderer Varietäten, insbesondere der ‚Standardvarietät' verknüpft sind.

Inkrementelle Funktionalität

Das klassische Wörterbuch ist eigentlich erst wirklich nutzbar, wenn es fertig ist. Das vor einem halben Jahrhundert begonnene Goethe-Wörterbuch, eine beeindruckende Leistung der deutschen Philologie, ist inzwischen beim Buchstaben M angelangt (http://germazope.uni-trier.de/Projects/GWB). Wenn man also nicht nur wissen will, was Goethe über *Gott* gesagt hat, sondern auch, was er über *Welt* gesagt hat, dann muss man warten. Dies schränkt den Nutzen merklich ein. Ein DLS hingegen geht schrittweise vor, und zwar so, dass in bestimmten Bereichen die Informationstiefe zunächst noch beschränkt ist; die Analyse durch den Lexikographen wird dann schrittweise fortgeführt. Bei einem Autorenwörterbuch beispielsweise wird man am Anfang einfach nur den Text in annotierter Form – d.h. mit gewissen Minimalinformationen versehen – zugänglich und über zweckmäßige Suchverfahren erschließbar machen. Dann ist noch nichts zu *Welt* gesagt, aber man kann sich die verschiedenen Verwendungen dieses Wortes in den verschiedenen Texttypen zusammenstellen und sich so selbst seine eigene Analyse erleichtern. Das ist vielleicht noch nicht so viel, wie man eigentlich will – aber es ist schon einmal hilfreich, wenn man über Goethes Begriff der Welt promovieren will.

Entsprechendes gilt natürlich, wenn das zugrundegelegte Korpus nicht auf einen Autor beschränkt ist. Man kann sich beispielsweise bei den syntaktischen Eigenschaften zunächst einmal mit den Ergebnissen einer ersten Analyse nach Wortklassen begnügen. Das ist sicher nicht, was man letztlich haben möchte. Es ist aber besser, als weitere Jahrzehnte zu warten, und für manche praktische Zwecke ist es vielleicht schon genug.

Kumulative Entwicklung

Damit ist das Gegenstück der inkrementellen Funktionalität auf Bearbeiterseite gemeint. Beim klassischen Wörterbuch erzwingt das Format eine Bearbeitung von A bis Z. Das wird umso problematischer, je reicher der abgebildete Ausschnitt des Lexikons sein soll. Bei einem DLS, das im Prinzip beliebig erweiterbar und in gewisser Weise nie abgeschlossen ist, können die einzelnen Bearbeitungsmodule zu verschiedenen Zeiten von verschiedenen Bearbeitern durchgeführt werden. Es muss lediglich eine zentrale Stelle geben, die das

Vorgehen koordiniert. So lassen sich viele konkrete Probleme der traditionellen Lexikographie mit ihren endlosen Arbeitszeiten weitgehend vermeiden.

Pluralität der Methoden

Das klassische Wörterbuch muss für jeden Typ von Eigenschaften einheitlich vorgehen: Man kann nicht für bestimmte Buchstaben mit einer ‚Head-driven Phrase Structure Grammar' und für andere mit traditioneller Schulgrammatik vorgehen. In einem DLS kann man hingegen je nach Zweck verschiedene Methoden miteinander verbinden. So kann man etwa für den gesamten Wortschatz mit den üblichen Bedeutungsumschreibungen, wie man sie aus dem traditionellen Wörterbuch kennt, beginnen und dann bestimmte Bereiche mit anderen Methoden differenzierter bearbeiten. Manche lexikalische Felder, beispielsweise die Modalverben, die Farbadjektive oder die zeitlichen Konjunktionen, sind so gut durchstrukturiert, dass hier eine Analyse über semantische Relationen wie Hyponymie, Antonymie usw. realistisch erscheint. Für andere Bereiche, insbesondere bei den Nomina, bietet sich ein solches Vorgehen nicht an. Es wäre ganz aussichtslos, etwa die verschiedenen Teile eines Automotors ausschließlich über Synonymie- oder Antonymierelationen beschreiben zu wollen. Hier versagt auch das Vorgehen des traditionellen Wörterbuchs weitgehend; es ist zwar nicht unmöglich, aber doch wenig hilfreich, die Bedeutung der Wörter *Pleuelstange* und *Kurbelwelle* durch eine Umschreibung in schlichten deutschen Worten dingfest zu machen. Viel hilfreicher sind hier Abbildungen, und so wird dies ja auch in Spezialwörterbüchern für technische Übersetzungen gemacht. In einem DLS können solche Bereiche jederzeit in eigenen Modulen über Bilder beschrieben werden.

All diese Charakteristika eines DLS sind letztlich nur verschiedene Ausprägungen der Flexibilität, die dadurch gewonnen wird, dass man sich vom Format des klassischen Wörterbuchs lösen kann. Das DWDS ist ein solches computerbasiertes Wörterbuchsystem. Es beruht auf Vorarbeiten, die an der Berlin-Brandenburgischen Akademie der Wissenschaften, ihren Vorgängerinstitutionen, aber auch an anderen wissenschaftlichen Einrichtungen geschaffen worden sind.

4 Ziele und Ausgangsbasis

Das Vorhaben hat zwei Hauptziele, die eng miteinander zusammenhängen:

A. Es soll das verfügbare lexikalische Wissen, wie es in den bisherigen großen Wörterbüchern seinen Niederschlag gefunden hat, zusammenführen und auf den neuesten Stand bringen.

B. Es soll ein digitales lexikalisches System entwickeln, das
   (a) Belege für die möglichen Verwendungen eines Wortes – in Form gut erschlossener Korpora – und eine wissenschaftlich verlässliche Beschreibung der verschiedenen Eigenschaften dieses Wortes miteinander verbindet,
   (b) sich jederzeit flexibel erweitern und korrigieren lässt, und
   (c) für viele – wissenschaftliche wie nichtwissenschaftliche – Zwecke nutzbar ist.

Ein solches Vorhaben lässt sich nur stufenweise verwirklichen. Vorgesehen sind drei Arbeitsphasen von jeweils sechs Jahren. Die drei Arbeitsphasen bauen aufeinander auf und bilden ein Ganzes. Sie sind jedoch so konzipiert, dass auch nach sechs bzw. zwölf Jahren ein Abschluss möglich wäre, ohne dass ein Torso von geringem Nutzen zurückbliebe. Anders als bei klassischen Wörterbuchvorhaben erfolgt die Bearbeitung nicht von A bis Z, sondern nach zunehmender Funktionstiefe: Das in der ersten Arbeitsphase avisierte Ergebnis hat bereits einen wissenschaftlichen und praktischen Nutzen, der erheblich über das, was wissenschaftliche oder kommerzielle Wörterbücher bisher leisten, hinausgeht; diese Arbeitsphase ist derzeit im Gang. In der zweiten und dritten Arbeitsphase, deren genaue Durchführung noch in der Planung ist, werden die Funktionen dann deutlich ausgebaut, vgl. dazu Abschnitt 5.

Ausgangsbasis sind vor allem die an der Berlin-Brandenburgischen Akademie der Wissenschaften (bzw. ihren Vorgängereinrichtungen) erarbeiteten Wörterbücher und Korpora. Dies sind:

A. Wörterbücher

1. *Deutsches Wörterbuch ([1]DWB)*. Das [1]DWB, von 1854 – 1960 in 32 Bänden veröffentlicht, umfasst etwa 330.000 Stichwörter. Eine digitale Version wurde am Trierer Kompetenzzentrum (http://germazope.uni-trier.de/ Projects/DWB) erstellt. Das [1]DWB ist nach wie vor das Hauptwerk der deutschen Lexikographie. Allerdings ist es durchweg überholt; dies gilt vor allem für die älteren Teile, die noch von Jacob und Wilhelm Grimm selbst bearbeitet wurden. Aber auch in den späteren Lieferungen wird der Wortschatz des 20./21. Jahrhunderts entweder überhaupt nicht oder doch nur unvollständig abgedeckt (Dückert 1987, Schmidt 2004).

2. *Deutsches Wörterbuch, Neubearbeitung ([2]DWB)*. Die Neubearbeitung der Buchstaben A bis F, vor fast fünfzig Jahren gemeinsam von der Göttinger und der Berliner Akademie begonnen, soll im Jahre 2013 abgeschlossen werden. Etwa die Hälfte der Lieferungen liegt in digitaler Form vor; eine Nachdigitalisierung der übrigen ist geplant (Wiegand 1989, Schmitt und Klein 2004).

3. *Wörterbuch der deutschen Gegenwartssprache (WDG)*. Das WDG, seit den Fünfzigerjahren an der Akademie der Wissenschaften der DDR erstellt und von 1964 bis 1977 in sechs Bänden veröffentlicht, enthält ungefähr 120.000 Stichwörter; etwa 3000 sind ‚DDR-lastig', allerdings meist nur in geringem Maße (Schaeder 1987, Wiegand 1990). Es ist in seinen Belegen überholt, besticht jedoch nach wie vor durch einen außerordentlich klaren Artikelaufbau und sehr gute Bedeutungsbeschreibungen. Das WDG wurde am Trierer Kompetenzzentrum digitalisiert und vom DWDS nach xml/tei-P5 strukturiert.

4. *Etymologisches Wörterbuch des Deutschen*. Das Etymologische Wörterbuch wurde in den Siebzigerjahren an der Akademie der Wissenschaften der DDR von einer Arbeitsgruppe unter Leitung von Wolfgang Pfeifer erstellt und 1989 in drei Bänden veröffentlicht. Es enthält Informationen zur Grammatik, Bedeutung und vor allem zur Wortgeschichte von circa 22.000 Lexemen. Diese sind in knapp 8.000 Haupteinträgen und ca. 14.000 Untereinträgen organisiert. Der diskursive Stil, den Pfeifer und seine Kollegen bei der Beschrei-

bung der Herkunft der Wörter verwenden, und die herausragende Qualität der Bedeutungsparaphrasen machen dieses Werk zu einem gut lesbaren und auch für den etymologischen Laien gut verständlichen und leicht zugänglichen Nachschlagewerk. Die mittlerweile durch die beiden BBAW-Projekte TELOTA und DWDS digitalisierte und aufbereitete Version des Wörterbuchs basiert auf der zweiten, 1993 im Akademie-Verlag erschienenen Auflage. Der Quelltext dieser Auflage wurde von Herrn Pfeifer noch einmal durchgesehen, korrigiert und ergänzt.

B. Textkorpora

Das ursprüngliche Textkorpus des DWDS – damals noch als Wörterbuch des 20. Jahrhunderts gedacht – wurde mit Unterstützung der Deutschen Forschungsgemeinschaft in den Jahren 2000-2003 erstellt; seither wird es kontinuierlich ausgebaut (Geyken 2007). Es setzt sich aus zwei großen Bestandteilen zusammen: dem kleineren, nach Textsorten ausgewogenen, öffentlich recherchierbaren Kernkorpus sowie einem wesentlich größeren Ergänzungskorpus; beide Korpora werden kontinuierlich ausgebaut.

Das Kernkorpus spiegelt den Wortschatz des gesamten 20. Jahrhunderts in größtmöglicher Ausgewogenheit wider. Bei der Erstellung des Kernkorpus wurde darauf geachtet, die Textsorten über das gesamte Jahrhundert gleichmäßig zu streuen und die prozentuale Verteilung der Textsorten untereinander angemessen zu berücksichtigen. Das Kernkorpus umfasst etwa 100 Millionen Textwörter. Aufgenommen wurden Dokumente aus fünf Bereichen:

(1) Schöne Literatur (unter Einschluss von Trivialliteratur, Kinderbüchern u.a.);
(2) Journalistische Prosa (Zeitungen, Magazine);
(3) Fachprosa (wissenschaftliche und populärwissenschaftliche Texte);
(4) Gebrauchstexte (Ratgeber, Verwaltungsvorschriften, Gebrauchsanweisungen, Theaterprogramme, Werbetexte);
(5) (Transkribierte) Texte gesprochener Sprache, (z.B. transkribierte Reportagen und Aufnahmen aus den Beständen des Deutschen Rundfunkarchiv Frankfurt/Babelsberg).

Bei der Auswahl der Texte im Kernkorpus wurde nicht nur auf gleichmäßige Streuung über den gesamten abgedeckten Zeitraum, sondern nach Möglichkeit auch auf Repräsentativität geachtet. Es erfasst daher den Sprachgebrauch des 20. Jahrhunderts zwar nach wie vor nicht vollkommen, aber besser als jedes andere Korpus. Derzeit wird das Kernkorpus sowohl in die Gegenwart als auch in die Vergangenheit erweitert. Für die 1. Dekade des 21. Jahrhunderts geschieht dies über Nutzungsvereinbarungen mit textgebenden Verlagen. Dieses Korpus befindet sich derzeit im Aufbau. Die Ausweitung der Textbasis auf Texte von vor 1900 bis ins Frühneuhochdeutsche findet im Rahmen des mit dem Digitalen Wörterbuch assoziierten Projekts *Deutsches Textarchiv* statt, das seit 2007 von der Deutschen Forschungsgemeinschaft gefördert wird (http://www.deutschestextarchiv.de). Die Prinzipien von Textauswahl, Textaufbereitung und Texterschließung sind dabei dieselben; ebenso wird dieselbe Suchmaschine verwendet.

Das Ergänzungskorpus umfasst vorwiegend neuere Zeitungstexte, da diese in großen Mengen elektronisch verfügbar sind. Unter anderem wurden dort größere Bestände folgender Zeitungen linguistisch erschließbar gemacht: *Berliner Zeitung, Bild, FAZ, Frankfurter Rundschau, Spiegel, SZ, Tagesspiegel, taz, Welt* und *ZEIT*. Aufgrund urheberrechtlicher Beschrän-

kungen sind diese Korpora teilweise nur intern verfügbar. Darüber hinaus wurden über Kooperationen auch kleinere Korpora zusammengestellt und elektronisch aufbereitet. Dazu zählt beispielsweise das C4-Korpus, welches Texte des 20. Jahrhundert aus Deutschland, Österreich der Schweiz und Südtirol enthält, ein Korpus ‚Jüdischer Periodika' des 19. und 20. Jahrhunderts sowie mehrere Korpora gesprochener Sprache. Das Ergänzungskorpus ist mit derzeit etwa einer Milliarde Textwörtern wesentlich umfangreicher; dabei handelt es sich jedoch im Wesentlichen um neuere Zeitungstexte, d.h. es ist nicht repräsentativ, aber für manche Zwecke von großem Nutzen – beispielsweise für statistische Untersuchungen, für Untersuchungen zur Mediensprache oder auch zur Veränderung der Orthografie in neuerer Zeit.

5 Arbeitsphasen

Das Vorhaben wird, wie schon bemerkt, in drei aufeinander aufbauenden, jedoch in sich abgeschlossenen Arbeitsphasen durchgeführt.

5.1 Phase A (1. 1. 2007 – 31. 12. 2012)

Diese Phase hat drei Teilziele:

- Vernetzung von Korpora und Wörterbüchern und damit der Integration des bisher verfügbaren lexikalischen Wissens,
- Entwicklung eines lexikographischen Arbeitsplatzes, der gleichermaßen für wissenschaftliche Zwecke wie für den gewöhnlichen Nutzer über Internet nutzbar ist,
- Ergänzung um eine Aussprachekomponente.

Was mit der ersten Teilaufgabe gemeint ist, lässt sich an einem einfachen Beispiel erläutern. Im [1] DWB wie im WDG – die hier exemplarisch stehen, für die anderen einbezogenen Wörterbücher gilt Entsprechendes – findet sich jeweils ein Eintrag für *absetzen*, der u.a. aus einer gegliederten Bedeutungsbeschreibung der verschiedenen Verwendungsweisen sowie aus einer Reihe von Belegen für diese Verwendungsweisen besteht. Im DWDS kann man sich diese beiden Einträge (ebenso weitere aus anderen Wörterbüchern) zunächst parallel in einem Fenster anzeigen lassen. Es ist möglich, alle Belege auszublenden, sodass nur noch die Struktur der Bedeutungsbeschreibung sichtbar und damit leicht vergleichbar ist. Ebenso ist es möglich, nur die Belege für einen bestimmten Zeitraum einzublenden, etwa nach 1800. So lässt sich leicht ermitteln, welche Wörter zu einer bestimmten Zeit ‚ausgestorben' (d.h. nicht mehr belegt) sind. Umgekehrt lassen sich auf diese Weise leicht Neologismen identifizieren. Bereits dies eröffnet der lexikologischen Analyse viele neue Möglichkeiten. Weiterhin sind die Artikel unmittelbar mit der Korpusdatenbank verbunden. Es lässt sich daher jederzeit die Verwendung eines Wortes in bestimmten Texten, Textgruppen oder auch bei bestimmten Autoren studieren. Dies macht es auch möglich, die vielfach exzellenten, aber nicht mehr aktuellen vorhandenen Analysen eines Wortes leichter und schneller zu überarbeiten und auf den neuesten Stand zu bringen.

Die dritte Teilaufgabe betrifft die größte Lücke in der lexikalischen Analyse der genannten Wörterbücher: Sie enthalten keine oder allenfalls marginale Angaben zu den lautlichen

Eigenschaften, wie denn die Aussprache ja allgemein ein Stiefkind der Lexikographie ist. Das hängt nicht zuletzt mit den Unzulänglichkeiten gängiger Lautschriften zusammen. Die technische Entwicklung hat es möglich gemacht, die einzelnen Wörter von ausgebildeten Sprechern sprechen zu lassen, sodass beispielsweise Sprachlerner ein realistisches Bild von der Aussprache erhalten. Grundlage ist – wie bei Siebs und dem Duden-Aussprachewörterbuch – die ‚deutsche Hochlautung', so wie sie beispielsweise von ausgebildeten Nachrichtensprechern verwendet wird. Es ist aber geplant, auch alle Wörter in anderen Varianten (insbesondere Wien und Zürich) aufzunehmen; dies ist aber nicht alleine Gegenstand des DWDS, sondern soll in Zusammenarbeit mit anderen Forschungsstätten geschehen.

Am Ende von Arbeitsphase A steht jedem Nutzer das gesamte verfügbare lexikalische Wissen, soweit es seinen Niederschlag in den einbezogenen Wörterbüchern gefunden hat, zur Verfügung, und dies in sehr leicht zugänglicher Weise. Dem Nutzer wird es ermöglicht, sein Wissen durch einen Sprung ins Korpus zu ergänzen, dort Verwendungen eines Wortes nachzusehen und damit gleichsam selbst zum Lexikographen zu werden; für manche Zwecke – etwa für einen literarischen Übersetzer, der nur nach einigen Belegen sucht – ist dies vielleicht schon hinreichend, für andere Zwecke nicht. Ebenso kann er sich die Aussprache eines Wortes anhören, eine Funktion, die vor allem für Sprachlerner von Nutzen ist. Am Ende dieser Arbeitsphase hat das DWDS daher schon einen klaren Wert für den Wissenschaftler wie für sonstige Nutzer, der wesentlich über alles Bisherige hinausgeht. Diese Arbeitsphase ist seit Anfang 2007 im Gang; der derzeitige Stand (Februar 2010) wird in Abschnitt 5 beschrieben.

Worüber das DWDS nach Abschluss der ersten Phase nicht verfügt, ist (a) eine semantische Analyse von Wörtern, die bislang in keinem der integrierten Wörterbücher erfasst sind, und (b) eine lexikalische Überarbeitung von Einträgen, die zwar vorhanden, aber fehlerhaft, unzulänglich oder einfach durch die Entwicklung der Sprache überholt sind. Die Arbeitsphasen B und C sind diesen beiden Aufgaben gewidmet. Anders als Phase A sind sind sie bislang nur in ihren Grundlinien geplant; ein – bisher nicht genanntes – Ziel der ersten Phase ist es, die Aufgaben der beiden weiteren Phasen zu konkretisieren.

5.2 Phase B (1.1. 2013 – 31. 12. 2018)

In dieser Phase sollen Wörter, die zwar in der Korpusdatenbank belegt, in den verknüpften Wörterbüchern aber nicht bearbeitet sind, lexikalisch analysiert werden. Diese Arbeit lässt sich aufgrund des in Phase A Erreichten wesentlich beschleunigen. Sie bedarf aber dennoch der Kompetenz ausgebildeter Lexikographen und ist entsprechend aufwändig. Für diese Phase sind daher zehn wissenschaftliche Mitarbeiter vorgesehen, davon zwei im computerlinguistischen Bereich und acht für die lexikalische Analyse.

Es wäre weder realistisch noch sinnvoll, alle im Korpus belegten, aber lexikalisch bislang nicht analysierten Wörter bearbeiten zu wollen. Das [1]DWB hat etwa 330.000 Stichwörter, das WDG etwa 120.000. In der derzeitigen DWDS-Wortdatenbank – die ja nur Texte des 20. Jahrhunderts enthält – finden sich hingegen etwa über vier Millionen Wörter. Die meisten davon kommen jedoch nur ein- oder zweimal vor, und viele sind nicht lexikalisierte Wortbildungen, die keiner eigenen lexikalischen Analyse bedürfen. Eine genaue Entscheidung, welche Wörter in dieser Phase bearbeitet werden sollen, ist noch nicht getroffen und lässt sich auch nicht treffen, bevor die Korpusdatenbank um ältere Texte erweitert worden

ist. Grundsätzlich gilt jedoch Folgendes. Das DWDS geht nicht, wie beim klassischen Wörterbuch, von A bis Z vor, sondern in zunehmender Tiefe der Bearbeitung. Es werden daher zunächst einmal all jene Wörter berücksichtigt, die (a) über einer gewissen Häufigkeitsschwelle liegen, und (b) die über verschiedene Texte streuen (also nicht beispielsweise nur bei einem Autor oder einem Texttyp belegt sind). Nach Abschluss dieser ersten Teilphase können nach Maßgabe der verfügbaren Zeit Häufigkeitsschwelle und Streuung geändert werden, sodass auch selten oder selektiv belegte Wörter bearbeitet werden können.

Nach dieser Phase sind häufig und von vielen verwendete neue Wörter beschrieben, andere, die selten oder nur von wenigen gebraucht werden, hingegen nicht. Hier ist der Benutzer wiederum auf die bestehenden Möglichkeiten verwiesen: Er kann sich jederzeit selbst Belege und Verwendungsweisen vor Augen führen.

5.3 Phase C (1.1. 2019 – 31. 12. 2025)

Am Ende von Phase B ist die lexikalische Analyse neuer Wörter auf dem neuesten Stand. Die meisten erfassten Wörter sind jedoch noch auf dem Stand von [1]DWB, (nur Buchstabe A bis F) oder WDG. In der letzten Arbeitsphase sollen diese Wörter aktualisiert werden. Auch diese Arbeit lässt sich erheblich schneller durchführen als bisher, aber wie bisher erfordert sie ausgebildete Lexikographen. Im Übrigen gilt das im vorigen Abschnitt Ausgeführte: Die Neubearbeitung soll nicht von A bis Z erfolgen, sondern nach Wichtigkeit in der Verwendung, so wie sie sich im Korpus widerspiegelt. Dabei mag es – wie schon in Phase B – durchaus sinnvoll sein, gelegentlich von rein statistischen Kriterien abzuweichen, beispielsweise weil ein Wort für eine bestimmte Epoche besonders wichtig ist, dann aber wieder verschwindet und deshalb insgesamt selten belegt ist. Die Gefahr, im Jahre 2025 aufzuhören und eine Baustelle zu hinterlassen, besteht nicht. Allenfalls ist die Analyse seltener Wörter nicht auf dem neuesten Stand; aber auch hier kann sich der Benutzer selbst aus den Quellen über die Verwendung eines Wortes, sein erstes oder letztes Vorkommen oder seinen Gebrauch in bestimmten Textsorten informieren.

6 Stand (Februar 2010)

6.1 Zusammenführung der Wörterbuch- und Korpusressourcen

Die Wörterbuch- und Korpusressourcen wurden beide gemäß den derzeitigen Empfehlungen der ‚Text Encoding Initiative' für Textkorpora und Wörterbücher (xml/tei P5) strukturiert. Die Korpora wurden darüber hinaus lemmatisiert und mit Wortartangaben versehen. Die Indizierung und Abfrage erfolgt über die linguistische Suchmaschine DDC (Dialing/DWDS-Concordancer), die auch lemmabasierte Abfragen bzw. die Suche nach Wortarten zulässt. Mit dieser Aufbereitung ist bereits eine einfache Verknüpfung von Stichwörtern der Wörterbücher und Korpora gewährleistet.

Eine erste weitergehende Vernetzung wurde für die beiden Wörterbücher, das auf dem WDG aufbauende DWDS-Wörterbuch und das Etymologische Wörterbuch, erarbeitet. Bei-

de wurden auf der Basis einer gemeinsamen ‚per Kopf' erstellten Metalemmaliste vernetzt, über die der Zugriff auf beide Wörterbücher gleichzeitig ermöglicht wird. Bei homographen Einträgen, wie z.B. *Abort* und *Pony*, mussten hierfür die jeweils zueinander passenden Artikel von Hand zugeordnet werden.

6.2 Überführung des WDG in das DWDS-Wörterbuch

Die zentrale Wörterbuchkomponente – das DWDS-Wörterbuch – beruht auf dem WDG (das in seinem historischen Bestand weiterhin zugänglich bleiben wird). Zunächst wurde der digitalisierte Text an die heutige Orthographie angepasst; Ziel war es, die von der seit August 2006 in Kraft getretenen Rechtschreibreform betroffenen Stichwörter des WDG zu identifizieren, zu markieren, ob diese Stichwörter gültig/ungültig sind und die Varianten zu diesen Stichwörtern zu verzeichnen. Ferner soll bei jedem von der Rechtschreibreform betroffenen Wort der Verweis auf das Regelwerk des Rats für deutsche Rechtschreibung erfolgen. Dieses Teilprojekt wurde 2007 begonnen und wird im 4. Quartal 2010 abgeschlossen sein. Weiterhin wurden mehr als 2000 ‚DDR-lastige' und heute nicht mehr vertretbare Definitionen und Beispiele des WDG von der Projektgruppe lexikographisch überarbeitet und in das DWDS-Wörterbuch eingetragen (wobei, wie bemerkt, die frühere Fassung weiterhin verfügbar ist).

Das WDG wurde in den Jahren 1961-1977 erarbeitet; es muss daher in den künftigen Arbeitsphasen nicht nur in seinem Stichwortbestand aktualisiert werden, sondern auch in seiner Binnenstruktur. Damit das effizient geschehen kann, war es notwendig, das Wörterbuch von einer xml/tei-Kodierung, die sehr stark an der typographischen Auszeichnung des WDG orientiert war, in eine systematisch inhaltsorientierte Datenbankstruktur zu überführen. In den vergangenen Jahren wurde dies von der Projektgruppe in mehreren Etappen geleistet (Schmidt et al. 2008), Herold und Geyken (2008). Beispiele für die Überführung sind die Vereinheitlichung von Deskriptoren (z.B. bei diasystematischen Angaben) oder die Rekonstruktion von Unterspezifizierung bei der Wortbildung. So kann beispielsweise aufgrund der unterspezifizierten Verwendung des Bindestrichs zu Markierung von Komposita-Reihen im Allgemeinen nur der menschliche Leser zwischen trennbaren und untrennbaren Verwendungen unterscheiden (*auf-, empor-, ent-, hervorsprießen* zu *sprießen*). Allgemein geht es somit darum, die Textverdichtungsmerkmale des gedruckten Wörterbuchs auf atomare Einheiten zurückzuführen. Insbesondere müssen hierbei viele Skopusprobleme ‚per Kopf' aufgelöst werden, die im gedruckten Wörterbuch implizit bleiben. Ein weiteres Beispiel hierfür ist die semantische Unterspezifizierung eines Separators wie des Kommas, welcher ohne Unterscheidung metasprachlich wie auch objektsprachlich gebraucht wird. Die Bestrebungen zur Textkompression im WDG führen darüber hinaus in vielen Fällen dazu, dass Bedeutungsparaphrasen im Bedeutungsteil des Eintrags nur fragmentarisch gegeben werden und die gesamte Bedeutung somit nur vom menschlichen Leser über den Rücksprung auf übergeordnete Lesarten oder durch Aufsuchen eines anderen Wörterbuchartikels erschlossen werden können (Herold/Geyken 2008).

6.3 Aussprachekomponente

Die Aussprachekomponente, die gemeinsam mit dem Max-Planck-Institut für Psycholinguistik (MPI-Nijmegen) entwickelt wird, wird in zwei Schritten erarbeitet. Im ersten Schritt werden von einer ausgebildeten Sprecherzieherin (Maren Böhm) alle im WDG lexikographisch bearbeiteten Stichwörter eingesprochen, durch das Institut für Sprechwissenschaft in Halle (Prof. Dr. Eberhard Stock und Prof. Dr. Ursula Hirschfeld) überprüft, gegebenenfalls korrigiert und dann als Audiodateien in die Wörterbuchartikelstruktur integriert. Auf der Webplattform können sie angeklickt und beliebig oft angehört werden. Im zweiten Schritt wird eine Auswahl aus den im WDG nicht enthaltenen, aber im DWDS-Korpus belegten hinzugenommen. Die wesentlichen Kriterien für diese Auswahl sind Grad der Lexikalisierung, Beleghäufigkeit und Streuung über Textsorten.

Derzeit stehen bereits 61.000 Audio-Files zum Abhören zur Verfügung; bis November 2010 sollen alle 90.000 Einträge freigeschaltet werden. Damit wird der erste Schritt abgeschlossen sein.

6.4 DWDS-Website

Das Konzept der DWDS-Website ist so angelegt, dass die verschiedenen Ressourcen, die das DWDS zur Verfügung stellt, je nach Recherchezweck von den Benutzern frei zusammengestellt und durchsuchbar sein sollen. Aus diesem Grund wurde die Website so konzipiert, dass jeweils eine Ressource in einem unabhängigen Fenster (einem so genannten ‚Panel') dargestellt wird und sich mehrere Fenster zu einer Sicht (‚View') zusammenfassen lassen (so genanntes ‚Panel-View'-Paradigma, siehe die Abbildung unten). Die Recherche erfolgt stets übergreifend in allen Fenstern einer Sicht. Derzeit stehen dem Benutzer etwa 20 verschiedene Ressourcen aus drei Bereichen zur Verfügung: Wörterbücher, Korpora und statistische Daten. Der *Wörterbuchbereich* enthält derzeit neben dem DWDS-Wörterbuch (sowie der elektronischen Variante der Druckausgabe des WDG) auch das ‚Etymologische Wörterbuch des Deutschen' . In Vorbereitung sind ferner Panels für das DWB (Erstausgabe und Neubearbeitung). Darüber hinaus können in diesem Bereich auch externe lexikalische Ressourcen eingebunden werden. Derzeit ist dies das als kollektives Projekt erarbeitete Synonymwörterbuch OpenThesaurus, geplant ist die Integration von GermaNet. Im *Korpusbereich* sind im öffentlich einsehbaren Bereich das Kernkorpus des 20. Jahrhunderts abfragbar, ferner auch das im Aufbau begriffene Korpus 21, das wöchentlich aktualisierte ZEIT-Korpus sowie mehrere andere Zeitungskorpora (Berliner Zeitung, BILD, Tagesspiegel). Darüber hinaus lassen sich *statistische Informationen* auf der Grundlage des Kernkorpus, der ZEIT und des BILD-Korpus visualisieren. Diese werden in der Form so genannter statistischer Wortprofile zur Verfügung gestellt und enthalten in Form von Wortwolken Wortverbindungen, die für ein gegebenes Suchwort statistisch gesehen besonders auffällig sind. Schließlich können für das Kernkorpus zeitliche Wortverlaufskurven mit genauen Zahlenangaben visualisiert werden.

Es gibt mehrere vorkonfigurierte Sichten:

- DWDS-Standardsicht (diese ist in der Abbildung gezeigt);
- Wörterbuchsicht,

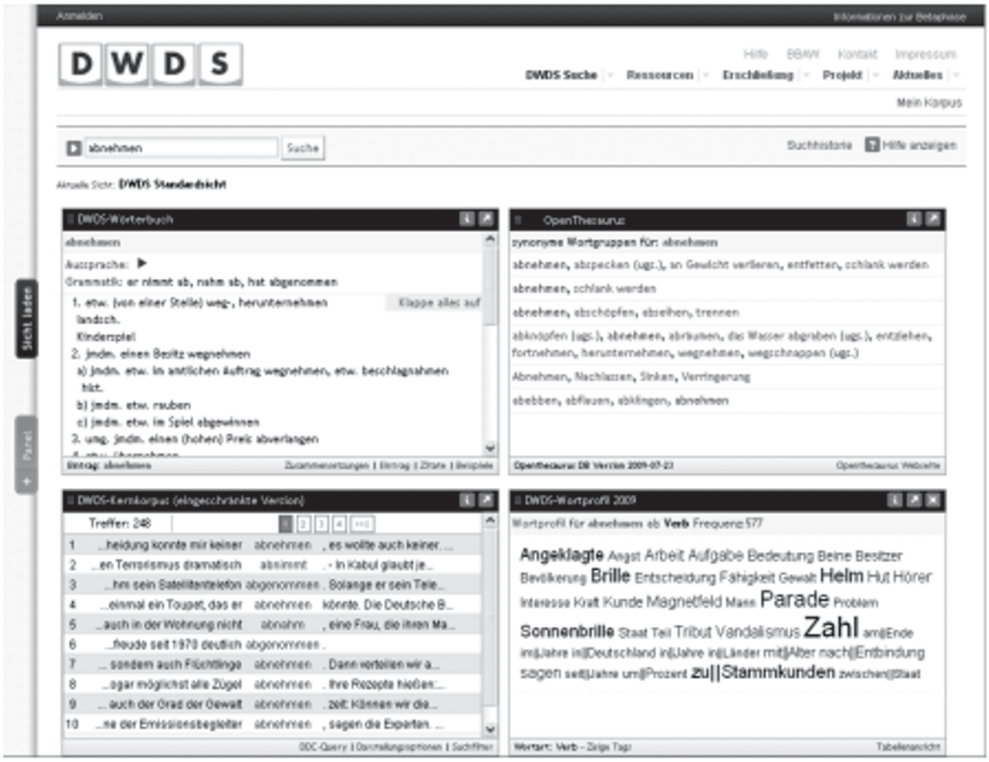

Abbildung: Auszug aus http://beta.dwds.de (Februar 2010)

– Korpussicht,
– Statistiksicht.

Voreingestellt für jeden Benutzer ist die DWDS-Standardsicht. Jeder Benutzer kann sich aber auch eigene Sichten aus den verschiedenen Panels selbst zusammenstellen; die Reihenfolge der Panels ist dabei frei konfigurierbar.

Nach der kostenlosen Anmeldung auf der Website kann ein Benutzer Belege in einem privaten Belegkorpus speichern. Jedem Beleg lassen sich individuelle Kategorien und optionale Kommentare zuweisen. So lassen sich je nach Bedarf eigene virtuelle Korpora bilden. Belegkorpora bilden insofern einen wichtigen Schritt auf dem Weg zu einem lexikografischen Arbeitsplatz, als diese – in einem künftigen noch zu implementierenden Schritt – den lexikografischen Zonen der Wörterbuchartikel zugeordnet werden können. Durch die direkte Kopplung des Belegkorpus mit den DWDS-Korpora wird gewährleistet, dass die gespeicherten Belege zu Definitionen in neu erstellten Wörterbucheinträgen stets auf vorhandene Ressourcen referenzieren. Es können auch aber auch eigene Korpora des Nutzers integriert werden.

Die neue Website basiert auf dem Django-Framework. Sie ist seit Juli 2009 als Beta-Version öffentlich zugänglich (http://beta.dwds.de) und wird im 2. Quartal 2010 die alte Website (http://www.dwds.de) ablösen Mit derzeit 27.000 angemeldeten Nutzern und durchschnittlich 200.000 Seitenaufrufen täglich gehört die DWDS-Website zu den drei wichtigsten akademischen Nachschlageportalen zur deutschen Sprache.

7 Schlussbemerkung

Der Sinn eines Wörterbuchs war und ist es, den Wortschatz einer Sprache zu bestimmten Zwecken zu beschreiben. Der Wortschatz einer Sprache ist aber nichts Statisches; er ist das, was in den Köpfen vieler Sprecher gestanden hat und steht, und er verändert sich fortwährend. Weder dieser Dynamik noch seiner schier unendlichen Komplexität kann ein gedrucktes Wörterbuch gerecht werden. Dazu benötigen wir ein Instrumentarium, das selbst in hohem Maße flexibel ist. Digitale lexikalische Systeme sind solche Instrumentarien – oder zumindest können sie es werden. Wilhelm von Humboldt schrieb 1810 über das Wesen der Arbeit an Universitäten und Akademien:

> Es ist ferner eine Eigenthümlichkeit der höheren wissenschaftlichen Anstalten, dass sie die Wissenschaft immer als ein noch nicht ganz aufgelöstes Problem behandeln und daher immer im Forschen bleiben [...]. (Humboldt 1906-1936, Band X, 251).

Wenn dies so ist, dann spiegeln Digitale Lexikalische Systeme das Wesen dieser Tätigkeit idealtypisch wider – nie abgeschlossen, aber auf jeder Stufe eine Bereicherung unseres Wissens und zugleich von Nutzen für viele.

Literatur

Dückert (Hg.) 1987 = Dückert, Joachim (Hg.): Das Grimmsche Wörterbuch. Untersuchungen zur lexikographischen Methodologie. Leipzig, 1987.

Geyken 2007 = Geyken, Alexander: The DWDS corpus: A reference corpus for the German language of the 20th century. Idioms and Collocations: Corpus-based Linguistic, Lexicographic Studies. London: Continuum Press. 2007.

Herold / Geyken 2008 = Herold, Axel / Geyken, Alexander: Adaptive word sense views for the dictionary database eWDG. The case of definition assignment. In: Storrer, Angelika, et al., Hrsg.: *Text resources and lexical knowledge. Selected papers from the 9th conference on natural language processing*. Berlin: Walter de Gruyter, 2008, S. 209-221.

Klein 2004 = Klein, Wolfgang. Vom Wörterbuch zum Digitalen Lexikalischen System. In: Zeitschrift für Literaturwissenschaft und Linguistik, Heft 136 , 2004, S. 72-100.

Murphy 2003 = Murphy, Lynne M.: Semantic relations and the lexicon. Cambridge [u.a.]: Cambridge University Press, 2003.

Pfeifer (u.a.) 1989 = Pfeifer, Wolfgang, u.a.: Etymologisches Wörterbuch des Deutschen. Berlin: Akademie-Verlag, 1989.

Schaeder 1987 = Schaeder, Burkhard: Germanistische Lexikographie. Tübingen: Niemeyer, 1987.

Schmidt 2004 = Schmidt, Hartmut: Das Deutsche Wörterbuch. Gebrauchsanleitung. In: Deutsches Wörterbuch. Elektronische Ausgabe der Erstbearbeitung von Jacob Grimm und Wilhelm Grimm,. Frankfurt/Main: Zweitausendeins 2004. S. 25-64. (DVD)

Schmidt (et al] 2008 = Schmidt, Thomas; Geyken, Alexander; Storrer, Angelika: Refining and exploiting the structural markup of the eWDG. Proceedings of EURALEX 2008, 15.-17, Barcelona, Juli 2008.

Schmitt/Klein 2004 = Schmitt, Peter; Klein, Wolfgang: Der alte und der neue Grimm. In: Die Brüder Grimm in Berlin. Katalog zur Ausstellung anläßlich des hundertfünfzigsten Jahrestages seit der Vollendung von Band 1 des Deutschen Wörterbuchs im Jahre 1854. S. Hirzel Verlag Stuttgart und Leipzig, 2004. S. 167-176.

Stackmann 2002 = Stackmann, Karl: Das Deutsche Wörterbuch als Akademieunternehmen. In: Die Wissenschaften in der Akademie. Vorträge beim Jubiläumskolloquium der Akademie der Wissenschaften zu Göttingen im Juni 2000, hrsg. von Rudolf Smend und Hans-Heinrich Voigt, Göttingen, 2002 . S. 247-319.

Wiegand 1989 = Wiegand, Herbert Ernst: Wörterbuchstile. Das Wörterbuch von Jacob Grimm und Wilhelm Grimm und seine Neubearbeitung im Vergleich. In: ders., Hrsg.: Wörterbücher in der Diskussion. Tübingen: Niemeyer, 1989, S. 227-278.

Wiegang 1990 = Wiegand, Herbert Ernst: Die deutsche Lexikographie der Gegenwart. In: Hausmann, Franz Josef, et al. (Hrsg.): Wörterbücher. Dictionaries. Dictionnaires. Ein internationales Handbuch zur Lexikographie […]. 2. Teilbd. (HSK 5.2). Berlin, New York 1990, S. 2100–2246.