Wolfgang Klein

Das digitale Wörterbuch der deutschen Sprache des 20. Jahrhunderts[1]

Ich möchte mich zunächst einmal sehr herzlich für die Einladung zu dieser Tagung bedanken. Ich finde es sehr beeindruckend, daß ein Verein etwas derartiges zustande bringt. Bedanken möchte ich mich auch bei Hartmut Schmidt für die freundlichen Worte, die mich - wie es im ‚Datterich' von Niebergall so schön heißt - schamrötlich machen und die ein großes Sprungbrett aufstellen, an dem gemessen der Sprung vielleicht etwas kurz ausfallen wird. Ich habe zwar vor längerer Zeit win wenig in der Computerlinguistik gearbeitet, auch im Zusammenhang mit Wörterbucharbeit, aber ich bin, wie Sie wissen, kein Lexikograph. Und wenn ich so all die bedeutenden Gelehrten hier sehe, die ich zum großen Teil nur dem Namen nach und aus ihren Veröffentlichungen kenne, da muß ich schon sagen, daß ich mir in diesem Kreise ein wenig vorkomme wie ein Dackel in einer Versammlung von Bernhardinern.

Das Projekt der Berlin-Brandenburgischen Akademie der Wissenschaften, über das ich berichten möchte, ist ein relativ großes Projekt - ein Projekt der hiesigen Akademie ja ist eigentlich immer eher bernhardinermäßig. Es knüpft an die lange Tradition lexikographischer Arbeit der Akademie an. Als man im Jahre 1700 dem damaligen brandenburgischen Kurfürsten Friedrich III vorgeschlagen hat, zu Berlin eine 'Societät der Wissenschaften' zu errichten, da wurde dieser Vorschlag nicht nur gnädigst angenommen, sondern der erlauchte Herr fügte von sich aus hinzu 'daß man auch auf die Kultur der deutschen Sprache bei dieser Fundation gedenken möchte, gleichwie in Frankreich eine eigene Akademie gestiftet'. Und so heißt es denn auch in der von Gottfried Wilhelm Leibniz entworfenen Stiftungsurkunde vom Juni desselben Jahres: 'soll bey dieser Societet unter anderen nützlichen Studien, was zur erhaltung der Teütschen Sprache in ihrer anständigen reinigkeit, auch zur ehre und zierde der Teütschen Nation gereichet, sonderlich mit besorget werden, also daß es eine Teütsch gesinnete Societet der Scientien seyn'. Was dem Kurfürsten vorschwebte, war vor allem ein großes Wörterbuch der deutschen Sprache; daher die Anspielung auf die AcadJmie franHaise. Damit hat sich die Societät und spätere Akademie Zeit gelassen. Erst zu Beginn des letzten Jahrhunderts hat sie das bedeutendste Werk zweier ihrer bedeutendsten Mitglieder unter ihre Fittiche genommen - Jacob und Wilhelm Grimms 'Deutsches Wörterbuch', das zu dieser Zeit schon mehr als ein halbes Jahrhundert in Arbeit war. Sie hat es dann, gemeinsam mit der Göttinger Akademie der Wissenschaften, bis zum Jahre 1960, also nach einem weiteren halben Jahrhundert, zum einem glücklichen Abschluß gebracht. Einige im Saal Anwesende waren ja direkt daran beteiligt. Eine lange Zeit, aber das Ergebnis ist sicher eine der größten lexikographischen Leistungen, die je vollbracht wurden. Aber es ist auch fair zu sagen, daß das DWB die

[1] Das folgende ist die etwas redigierte, von Versprechern und anderen Errata befreite Nachschrift eines freien Vortrags; sie hat daher noch den Charakter eines solchen Vortrags. Kerstin Mauth hat mir bei der Überarbeitung sehr geholfen. Ihr und Jürgen Scharnhorst danke ich sehr für ihre Unterstützung.

Sprache des 20. Jahrhunderts nur sehr unzulänglich erfaßt. Es gibt zwar einige Belege aus den ersten Jahrzehnten dieser Zeit, und manche von ihnen möchte man heute gar nicht so gerne erwähnen, aber insgesamt ist dieser Zeitraum sicherlich nicht abgedeckt. Eine Neubearbeitung der Buchstaben A-F ist, wie Sie wissen, unterwegs, und sie wird hoffentlich in nicht allzulanger Zeit fertig sein. Aber einige Jahre wird schon noch dauern, und auch dann werden nur die Buchstaben A-F auf neuerem Stand sein allerdings auch nicht auf dem neuesten, denn seit Erscheinen der ersten Lieferungen der Neubearbeitung sind auch schon wieder einige Jahre ins Land gegangen.
Die Akademie hat dann Mitte der fünfziger Jahre nicht nur einen großen Anlauf zu einem Wörterbuch der Gegenwartssprache gemacht, sondern es in relativ kurzer Zeit auch zum Abschluß gebracht. Ich selbst als ein Nichtbeteiligter finde, daß das WDG eine außerordentliche Leistung ist. Das sage ich jetzt nicht als bloße Schmeichelei für einige hier im Saal Anwesende, die zu seinen Mitarbeitern zählen! Vor allen Dingen die Klarheit der Artikel ist immer wieder beeindruckend, wenn man sie mit anderen Wörterbüchern der deutschen Sprache und auch anderer Sprachen vergleicht. Aber auf der anderen Seite muß man sicherlich feststellen, daß das WDG nicht die deutsche Sprache des 20. Jahrhunderts in ihrer Breite abdeckt. Das kann zum Teil schon mal aus historischen Gründen gar nicht sein, weil das letzte Drittel dieses Zeitraums naturgemäß nicht mehr erfaßt werden konnte. Aber auch die Auswahl der Quellen ingesamt ist, - wie es, glaube ich, heißt - an der Sprache der ‚bildungstragenden Schicht' orientiert; es gibt viele Bereiche der deutschen Sprache, die dadurch mehr oder minder ausgeschlossen sind. Ich habe mir im Zusammenhang mit einer von uns geplanten Digitalisierung des WDG noch einmal einige Artikel angeschaut, und es ist schon so, daß - zumindest in den ersten Teilen – Wilhelm Raabe mehr zitiert wird als beispielsweise Bert Brecht oder Anna Seghers. Dies gilt im übrigen auch im Vergleich mit dem Dudenwörterbuch, das ja nicht zuletzt vom WDG inspiriert wurde und das in seiner neuesten, zehnbändigen Ausgabe wirklich ein beeindruckendes Werk ist, von dem man aber auch schwerlich sagen kann, daß es eine vollständige lexikographische Abdeckung der Sprache des 20. Jahrhunderts bietet. Wenn Sie sich das Quellenverzeichnis ansehen, so liegt das Schwergewicht eindeutig auf literarischer Sprache; viele andere Bereiche werden nicht oder kaum berücksichtigt. Auch bei der literarischen Sprache ist im wesentlichen an die ‚hohe Literatur' gedacht, die sogenannte Trivialliteratur wird kaum erfaßt. Auch das ist nicht so sehr als Kritik gemeint, das ‚Dudenwörterbuch' ist ohne jeden Zweifel eine bemerkenswerte Leistung. Aber man kann wirklich nicht behaupten, daß es derzeit eine systematische und erschöpfende Darstellung des deutschen Wortschatzes des 20. Jahrhunderts in seiner ganzen Breite gäbe. Kein Kompliment für eine Kulturnation.
Die Berlin-Brandenburgische Akademie hat sich vor einigen Jahren an ein solches Vorhaben gemacht, der erste Vorschlag dazu ist von Hartmut Schmidt gekommen. Vor etwas mehr als vier Jahren ist eine kleine interne Kommission der BBAW eingesetzt worden, die sich dies einmal überlegen sollte. Dieser Kommission haben Manfred Bierwisch, Hartmut Schmidt, Dieter Simon - der Präsident der Akademie - und ich selbst als Leiter angehört. Wir haben dort den Plan gefaßt, ein wirklich neues Wörterbuch der deutschen Sprache des 20. Jahrhunderts zu schaffen, und zwar ein Wörterbuch, das durch folgende Eigenschaften gekennzeichnet ist.

Erstens sollte es wirklich die gesamte Sprache des 20. Jahrhunderts abdecken, soweit daß überhaupt irgendwie möglich ist. Zweitens sollte es nicht beschränkt oder auch nur

konzentriert sein auf die Gegenwartssprache, sondern es sollte auch die deutsche Sprache um 1900 oder die Sprache zur Hitlerzeit abdecken. Drittens: es sollte - was Textsorten anbetrifft - relativ breit sein, also natürlich die literarische Sprache berücksichtigen, aber andere Verwendungen gleichberechtigt einbeziehen. Viertens: es sollte für einen breiten Nutzerkreis geeignet sein, d.h. für Wissenschaftler unterschiedlicher Disziplinen, aber durchaus auch für Journalisten, Übersetzer, überhaupt für alle, die sich aus dem einen oder andern Grund für die deutsche Sprache interessieren. Der fünfte und wahrscheinlich wichtigste Punkt ist folgender: das neue Wörterbuch sollte so weit als möglich mit den Mitteln der Computerlexikographie erstellt werden. Das Ziel war also nicht direkt ein gedrucktes Wörterbuch - seien es nun drei oder zwanzig Bände. Es sollte vielmehr primär ein ‚lexikographisches System' auf dem Computer erstellt werden, durchaus mit der Vorstellung, daß man hinterher einzelne Papierwörterbücher daraus ableiten könnte. Dies war jedoch nicht das primäre, sondern eher ein nachgeordnetes Ziel.

Wir haben uns dann weiterhin überlegt, ob wir das überhaupt schaffen können. Sie wissen, daß es bei den heutigen finanziellen Gegebenheiten auch einer Akademie vollkommen unmöglich ist, ein Projekt auf die Beine zu stellen, an dem - sagen wir - zehn Mitarbeiter hauptamtlich beteiligt sind. Die Zeiten sind leider lange vorüber. Das war auch einer der Gründe dafür, weshalb wir das Vorhaben von Anfang an als ‚digitales Wörterbuch' anlegen wollten: auf diese Weise hat man eine erheblich höhere Flexibilität als bei einem von A – Z angelegten gedruckten Wörterbuch; ich komme darauf noch ausführlich zurück. Wir haben dann beschlossen, zunächst einmal eine - wie man heute so schön sagt - , ‚Machbarkeitsstudie' zu machen (ich weiß nicht ob dieses Wort schon in irgendeinem Wörterbuch steht).

Die erste eigentliche Arbeitsphase bestand dann darin, eine sogenannte ‚Demo-CD' zu erstellen, das heißt eine kleine Corpussammlung von ungefähr zehn Millionen Wörtern mit einer geeigneten, relativ intelligenten Software, die auf einer CD Platz hat. Diese Demo-CD ist auch im Internet verfügbar; Sie können Sie sich unter der Webadresse **www.dwds.de** ansehen. Auf dieser Grundlage haben wir uns nun die Finanzierungsmöglichkeiten überlegt. Der nächste Schritt bestand darin, einen relativ umfangreichen Antrag an die DFG (etwa 1,8 Mio DM) zu stellen. Er wurde zu unserer Freunde auch bewilligt, sodaß wir im März 2000 mit der konkreten Arbeit beginnen konnten. Dafür sind wir der DFG und ihren Gutachtern außerordentlich dankbar, wie man denn diese wunderbare Institution gar nicht genug loben kann.

Wo ich schon beim Loben bin - jetzt ist vielleicht der Zeitpunkt gekommen, etwas über die Personen zu sagen, die an dem Vorhaben beteiligt sind. Da ist außer den vorhin schon genannten Bierwisch, Schmidt, Simon und Klein eine Reihe von Experten, die wir bei wechselnden Gelegenheiten hinzugeladen haben, die uns sehr engagiert geholfen haben und die ich hier nicht einzeln aufzählen kann; unser Dank ist ihnen aber gewiß. Und dann haben wir natürlich die Leute, die die eigentliche Arbeit machen: es ist ja immer so, daß es da einen gewissen Unterschied gibt zwischen denen, die sagen, was gemacht werden soll, und denen, die es wirklich machen (‚Wer baute das siebentorige Theben?'). Das sind in erster Linie Ralf Wolz, der für die Textbeschaffung zuständig ist; zweitens Gerald Neumann, Computerexperte und Lexikograph, der auch hier anwesend ist, und als Leiter des Ganzen Alexander Geyken, Philologe und Computerlinguist. Wir haben außerdem ein relativ erlauchtes Kuratorium, das das Ganze unter seinen Schirm genommen hat und dem eigentlich durch Zufall als Vorsitzender unser Bundespräsident angehört, also

Johannes Rau, aber auch Richard von Weizsäcker, der frühere Bundespräsident, weiterhin Wolfgang Frühwald, Hans Magnus Enzensberger, Uwe Honnefelder, Wolf Lepenies, Christian Meier und Dieter Zimmer. Die meisten von diesen sind naturgemäß nicht übermäßig engagiert, haben uns aber sehr, sehr viele Kontakte zu Verlagen und anderen Institutionen eingebracht, und insbesondere Hans Magnus Enzensberger, ein großer Wörterbuchliebhaber, hat zahlreiche konkrete Vorschläge gemacht und einzelne Texte für uns geschrieben.

Die eigentliche Arbeit am DWDS besteht hier - wie praktisch bei allen Wörterbüchern - aus zwei großen Phasen. Wenn ein Wörterbuch nicht primär aus der Exzerption anderer Wörterbücher lebt, sondern aus den Quellen gearbeitet ist, muß das einfach so sein: Es gibt zunächst eine erste Phase, die der Corpuserstellung oder der Sammlung der Belege dient; die zweite Phase ist die der eigentlichen lexikographischen Ausarbeitung. Ich habe ‚Phasen' gesagt, obwohl das – wie jeder, der es mal versucht hat, weiß - eigentlich nicht richtig ist. Es kann auf keinen Fall so sein, daß man zunächst das Eine abschließt und dann das Andere anfängt, sondern die Arbeiten überlappen sich fortwährend. Das hat schon für die traditionelle Wörterbucharbeit gegolten, wo man allerdings in seiner Flexibilität beschränkt ist; wenn die ersten Lieferungen erschienen sind, dann kann man nicht noch nachträglich Belege einfügen, es sei denn, es gibt – wie beim WDG - neue Auflagen. Beim Computerwörterbuch oder im Computersystem, einem lexikalischen System auf dem Computer also, ist man da wesentlich flexibler, weil man das Corpus jederzeit ergänzen und neue Belege relativ mühelos einarbeiten kann. Daher kann man noch weniger von zwei getrennten Phasen sprechen; trotzdem ist es praktisch, zunächst einmal einfach der Klarheit der Darstellung halber diese Redeweise beizubehalten. Was wir also zuerst in Angriff wollten, ist das Corpus. Dazu werde ich jetzt gleich Näheres ausführen, bevor ich dann ich dann auf die Planung der lexikographischen Ausarbeitung zu sprechen komme; zum Abschluß werde ich auf einen ganz konkreten Teil der lexikographischen Ausarbeitung etwas näher eingehen.

Zunächst also zum Corpus. Da hat man natürlich immer das große Problem, daß es so etwas wie ein wirklich ‚repräsentatives' Corpus - und sei es auch noch so groß - im Grunde nicht geben kann. Die Sprache ist zu reich in ihren vielfältigen Erscheinungsformen, und in vielen Fällen hat man gar keinen oder nur einen sehr schwierigen Zugang zu wesentlichen lexikalischen Quellen. Wir haben uns entschlossen, zwei Corpora aufzubauen. Das eine nennen wir das ‚Kerncorpus'. Dies ist jener Teil, an dem uns eigentlich gelegen ist. Es ist zunächst einmal in seinem Umfang auf ungefähr hundert Millionen Wörter laufenden Textes begrenzt, soll sich aber wirklich gleichmäßig über das gesamte zwanzigste Jahrhundert erstrecken - vom 1900 bis 1999 bis zum Ende. Das zweite Corpus nennen wir ‚Ergänzungscorpus'; es ist dies ein opportunistisches Corpus, d.h. es sind Texte, die wir nicht nach wohlüberlegten Kriterien aussuchen, sondern die man mehr oder minder nimmt, wo und wie man sie bekommt. Gelegenheit macht Texte. Es gibt ja sehr viele Texte, die bereits digital verfügbar sind, vor allen Dingen neuere Zeitungsausgaben, und die kann man im großen Maßstab kaufen; nicht selten sind sie auch umsonst zu haben. Dies ist ja auch an verschiedenen Stellen bereits in der einen oder anderen Form geschehen. An einem solchen opportunistischen Corpus ist uns nicht primär gelegen. Für viele statistische Zwecke und für selten belegte Wörter ist es aber sicher sehr nützlich, und so haben wir uns eben entschlossen, auch Texte dieser Art zu sammeln. Derzeit beläuft sich unser Ergänzungscorpus übrschlagsweise auf

fünfhundert Millionen Wörter. Aber unser eigentliches Ziel gilt dem Kerncorpus, das, wie gesagt, zunächst hundert Millionen Wörter umfassen soll und dessen Zusammensetzung ich Ihnen jetzt zumindest kurz erklären will.

Es gibt verschiedene Aspekte, die man da berücksichtigen muß: erstens, die *Zusammensetzung*, zweitens die Art der *Erfassung*, d.h. wie man eigentlich die Daten in digitaler Form bekommt, und drittens die *Annotation*, d.h. wie man die Daten aufbereitet, so daß sie wirklich für den Lexikographen sinnvoll nutzbar sind. Zunächst also zur Zusammensetzung. Unsere Vorstellung war, wie schon gesagt, sehr unterschiedliche Texttypen zu berücksichtigen und das gesamte 20. Jahrhundert abzudecken. Vielleicht sollte ich hier noch einmal kurz einfügen, daß das 20. Jahrhundert vom Lexikographischen her ein außerordentlich reicher Zeitraum ist, wenn man das Spektrum dessen, was in diesem Zeit alles geschehen ist und geschrieben wurde, betrachtet. Die *Buddenbrocks* sind vor ziemlich genau hundert Jahren erschienen. Wenn Sie mal überlegen: das ist neunundsechzig Jahre nach Goethe's Tod im selben Jahr, in dem auch *Faust II* erschienen ist. Es sind dreiundneunzig Jahre, nachdem *Faust I* erschienen ist. Die *Buddenbrocks* erschienen weniger als vierzig Jahre nach Stifters *Witiko*; das ist weniger, als uns vom ersten Erscheinen der *Blechtrommel* trennt (und, nebenbei bemerkt, länger als die ganze deutsche Romantik) . In einem Satz: Thomas Mann, den wir als einen großen Autor der 20. Jahrhunderts empfinden, steht mit seinem berühmtesten Werk zeitlich – und vielleicht auch sprachlich - gesehen einfach näher an Autoren wie Stifter, Raabe oder selbst Goethe als an uns.

Unsere Vorstellung war also, das gesamte Spektrum abdecken zu wollen. Das bezieht sich erstens auf den Zeitraum: es sollten in gleichmäßiger Verteilung Belege von 1900 bis 1999 erhoben werden. Warum ausgerechnet dieser Zeitraum? Sollte man nicht eher 1918 anfangen? Oder 1914? Oder 1933? Letztlich ist jede solche Abgrenzung nicht ohne Willkür, und da scheint es uns sauberer, nicht einen inhaltlichen Grund vortäuschen zu wollen, sondern dazu zu stehen: es ist einfach das vergangene Jahrhundert. Allenfalls soll das Belegmaterial in die ja stetig fortschreitende Gegenwart weitergeführt werden. Nichts spricht im Prinzip dagegen, zu einer späteren Zeit eine ähnliche Verlängerung in die Vergangenheit vorzunehmen: ein Computercorpus ist sehr flexibel.

Zweitens sollten auch unterschiedliche Textsorten in einigermaßen gleichmäßiger Verteilung berücksichtigt werden. Wie Sie sicher wissen, gibt es eine endlose Diskussion darüber, welche Textsorten man nach welchen Kriterien unterscheiden soll; wie immer man sich hier entscheidet, es wird stets andere Auffassungen geben, und das nicht ohne nachvollziehbare Gründe. Was wir hier machen, ist etwas relativ Praktisches. Wir haben uns auf fünf größere Typen festgelegt: erstens, Belletristik; zweitens, journalistische Prosa; drittens, wissenschaftliche Prosa; viertens, Gebrauchstexte; und fünftens, gesprochene Sprache. Und im Geiste höre ich schon einige sagen: „Ja, sind diese Klassifizierungen überhaupt berechtigt? Kann man das nicht anders machen?“ Ja, das kann man absolut anders machen, aber für uns ist das eigentlich primär eine praktische Einteilung, da nämlich jedes einzelne Dokument und damit auch jeder einzelne Beleg genau klassifiziert ist und sich jederzeit erschließen läßt. Man kann also, wann immer man will, die Klassifizierung feiner machen oder aber auch irgendwie zusammenschrumpfen lassen.

Die Zusammensetzung ist ungefähr die folgende: wir haben uns entschieden, ungefähr fünfundzwanzig Prozent - also ein Viertel - Belletristik zu nehmen; und zwar nicht nur

hohe Literatur, sondern durchaus auch Trivialliteratur. Weitere fünfundzwanzig Prozent sind journalistische Prosa – Zeitungstexte, Wochenschriften usw. Das macht zusammen also die Hälfte des Belegmaterials aus. Dann haben wir zwanzig Prozent Fachprosa, d.h. wissenschaftliche Texte, wobei ‚wissenschaftlich' nicht unbedingt im Sinne von wissenschaftlichen Originalveröffentlichen zu verstehen ist; wir haben auch zu einem grossen Teil Arbeiten aus der Akademie-Zeitschrift "Forschungen & Fortschritt" ausgenutzt, in der die bedeutensten Autoren ihrer Zeit geschrieben haben - Wissenschaftler wie Köhler, Einstein, Planck oder beispielsweise Friedrich Maurer für die Germanistik. Diese Aufsätze sind durchaus wissenschaftlich, tendieren aber schon ein wenig ins Allgemeinverständliche. Zu diesen bis jetzt insgesamt siebzig Prozent kommen zwanzig Prozent Gebrauchstexte hinzu, die wiederum ein weites Spektrum bilden. Darunter fallen zum Teil Rechtstexte, juristische Texte, Gesetze oder auch Verordnungen, aber auch so etwas wie Kochbücher, Tanzlehrgänge und dergleichen, oder auch Anweisungen zum Autoreparieren. Dadurch entsteht also eine relativ weite Bandbreite von lexikographisch verwertbarem Material. Die letzten verbleibenden zehn Prozent ergeben sich aus der gesprochenen Sprache.

Wenn ich jetzt etwas näher auf die Auswahl eingehe, lassen Sie mich gleich mit der gesprochenen Sprache anfangen, weil das im Moment unsere Schwachstelle ist. Es ist nicht allzu schwer, transkribierte Texte gesprochener Sprache aus jüngster Zeit zu bekommen, - gerade heute habe ich nochmal mit einer früheren Mitarbeiterin von mir drüber verhandelt, die ein sehr schönes, großes Corpus solcher Texte hat, das auch transkribiert ist. Was aber naturgemäß sehr selten zu finden ist, sind Texte gesprochener Sprache aus der ersten Jahrhunderthälfte. Vergangen wie Schall und Rauch. Wir sind sehr stolz, daß es uns gelungen ist, durch eine Zusammenarbeit mit dem Deutschen Rundfunkarchiv in Babelsberg eine Reihe von Hörfunkreportagen aus den zwanziger und dreißiger Jahren zu bekommen. Das Deutsche Rundfunkarchiv reicht von 1923 an durch das gesamte Jahrhundert hindurch, so daß wir Zugang zu einer doch vergleichsweise repräsentativen Auswahl von Texten gesprochener Sprache haben. Die müssen aber transkribiert und natürlich in den Computer aufgenommen werden - und da hinken wir im Moment deutlich hinterher; möglicherweise müssen wir uns hier mit weniger als den geplanten 10% am Gesamtmaterial bescheiden; aber all dies sind ohnehin nur ungefähre Werte.

Lassen Sie mich jetzt zu den anderen vier Bereichen kommen, zunächst zur Belletristik. Die Idee ist, daß wir zunächst für jedes Jahr des gesamten Jahrhunderts im Durchschnitt ungefähr vier größere Prosawerke aufnehmen - wobei jeweils noch einige Gedichte und Dramen dazukommen. Ich hab mir mal angesehen, was wir beispielsweise für das Jahr 1930 haben. Da haben wir Penzoldts *Die Powenzbande* - ein sehr schönes Buch, ganz am Rande bemerkt. Wir haben Ina Seidels *Wunschkind*, was wahrscheinlich nicht viele gelesen haben (es lohnt sich eigentlich auch nicht), wir haben dann Fritz Steubens *Tecumseh*, an das sich wahrscheinlich viele aus ihrer Kindheit erinnern werden (ein sehr schönes Jugendbuch), und schließlich haben wir den *Mann ohne Eigenschaften*, den natürlich alle gelesen haben hier, d.h. den ersten Band, der 1930 erschienen ist. Wie gesagt, die genaue Zahl schwankt ein bißchen, aber im Schnitt sind es vier größere Werke pro Jahr. Sie können sich den größten Teil dieser Auswahl inzwischen unter der eben genannten Internetadresse - also **dwds.de –** ansehen. Da steht unter dem Stichwort

"Matrix" die jetzige Auswahl; sie wird sich vielleicht noch ein wenig ändern; aber im wesentlichen wird es dabei bleiben.
Es liegt immer ein gewisses Moment der Willkür darin, welche Texte man in der Tat aufnimmt. Wir alle haben hier unsere Vorlieben, und meine kleinen Kommentare eben zeigen, wie beispielsweise die meinen sind; ich finde schon, daß man Werner Bergengruen und Leo Perutz aufnehmen muß. Wir sind, um dies zumindest ein wenig zu objektivieren, so vorgegangen, daß wir in unserer Arbeitsgruppe zunächst einmal eine Vorauswahl getroffen haben. Dann haben wir - da es nun schon mal um ein Akademiewörterbuch geht - einfach die Expertise unserer Akademie genutzt; schließlich zählt sie viele der bekanntesten Wissenschaftler der Republik zu ihren Mitgliedern. Wir haben also die Literaturwissenschaftler (Wolfgang Frühwald, Wilhelm Vosskamp und Conrad Wiedemann) in der BBAW gefragt und sie gebeten, für jedes Jahr des vergangenen Jahrhunderts die Werke anzugeben, die sie für die wichtigsten und prägendsten halten. Aufgrund dieser Auskünfte haben wir unsere provisorische Liste revidiert, so daß wir sagen können, daß wir uns nicht nur auf unser Urteil stützen, sondern durchaus auch das von anderen angesehenen Experten berücksichtigen. Daßelbe Verfahren haben wir dann für den dritten Bereich, die Fachprosa, angewandt (zu dem zweiten Bereich, Zeitungstexte, sage ich gleich noch etwas). Wir haben wissenschaftliche Texte, auch gleichmäßig über das ganze Jahrhundert, gestreut aufgenommen, und hier haben wir es einfach so gemacht, daß wir an die Akademiemitglieder per email einen Rundbrief geschickt und sie gebeten haben, für jedes Jahrzehnt drei oder vier der bedeutenden Werke zu nennen, die nach ihrer Einschätzung die entscheidenen Werke in ihrer Disziplin waren. Da ist etwas in mancher Beziehung sehr Trauriges herausgekommen, traurig jedenfalls für jeden, der die deutsche Sprache liebt: es ist nämlich so, daß vielleicht ab 1980 die bedeutenden wissenschaftlichen Werke nicht mehr auf Deutsch erscheinen, wenn man von einigen wenigen Fächern absieht. Es gibt also sehr wenig, was man da anführen könnte, allenfalls populärwissenschaftliche Veröffentlichungen. Aber in früheren Zeiten gab es natürlich viele deutsche Publikationen von Rang, und es hat uns sehr erstaunt, welche Resonanz diese Befragung an der Akademie hatte. Viele haben sich zum ersten Mal überlegt, was denn wirklich die prägenden Werke ihrer Disziplin sind: Geschichte, Mathematik, Psychologie, Ingenieurwissenschaften usw. Daraus haben wir jetzt eine Auswahl getroffen, die über das Jahrhundert gestreut ungefähr zwanzig Prozent unseres gesamtes Corpus ausmacht.
Jetzt komme ich nochmal zurück zu den Zeitungstexten. Da gibt es wiederum einen Bereich, der sehr leicht zu bearbeiten ist, und das sind die neueren Texte. Aber das Schwierige sind die Zeitungen - meistens noch in Fraktur gedruckt - aus der ersten Jahrhunderthälfte. Wir haben uns einige der bedeutensten Zeitung vorgenommen, und zwar solche, die möglichst lange erschienen sind - ganz durchgängige gibt es unsres Wissens nicht - , im wesentlichen aus Berlin, Köln und München. Es gibt auch noch solche aus Österreich und der Schweiz. Wir haben also das *Berliner Tageblatt* von Anfang an, also seit 1900, aufgenommen. Wir haben die *Vossische Zeitung*, so lange sie erschienen ist, erfaßt. Darüber hinaus sind eine Münchner Zeitung und eine Kölner Zeitung aufgenommen worden. Wir haben zu späteren Zeiten versucht zu balancieren; d.h. wir haben beispielsweise aus der Zeit des Dritten Reiches sowohl den *Völkischen Beobachter* (was man einfach aus lexikographischen Gründen machen muß), aber auch eine Reihe von Exilzeitungen berücksichtigt, so daß hier eine gewisses Gleichgewicht

gewährleistet ist. Das geht also durch das ganze Jahrhundert durch, und zwar in folgender Weise. Zunächst wurden für jedes Jahr fünf bis sechs vollständige Ausgaben einer der genannten Zeitungen aufgenommen. Wir haben aber außerdem noch jeweils eine Auswahl von Artikeln hinzugenommen, die sich um bedeutende Ereignisse in dem betreffenden Jahr ranken. Das war übrigens eine der Ideen, die Hartmut Schmidt in die Diskussion gebracht hat. Wir haben z.B. im Jahre 1930, dem Jahr, zu dem ich Ihnen vorhin auch die literarischen Texte genannt habe, Ereignisse wie die Erstaufführung des "Blauen Engel" dokumentiert, dann, daß Max Schmeling Weltmeister im Schwergewicht geworden ist, oder auch, daß zum ersten Mal ein Nazi, nämlich Frick, in eine Regierung eingetreten ist. Zu solchen Ereignissen, über die normalerweise alle Zeitungen berichten, haben wir dann eben aus allen Zeitungen die entsprechende Berichterstattung aufgenommen. In manchen Fällen ist es so, daß einschlägige Berichte sowohl in den politischen Seiten als auch im Feuilleton auftauchen. Daß Schmeling Weltmeister gegen Jack Sharkey geworden ist, wurde, wie man sich vorstellen kann, in verschiedenen Zusammenhängen aufgegriffen, und entsprechend ist der Wortschatz vielleicht auch ein bißchen anders. Auf diese Art und Weise denken wir ein wenig dem Problem begegnen zu können, das man dadurch hat, daß Zeitungen oder besser, journalistische Prosa, keine sehr einheitliche Gattung ist, sondern die Sprache erheblich variiert, je nachdem in welchem Teil der Zeitung der Bericht steht.

Soviel zu den drei Kategorien journalistische Prosa, Belletristik und Fachprosa. Unsere dritte Kategorie sind Gebrauchstexte. Auch da haben wir versucht, nach einigermaßen ausbalancierten Gesichtspunkten vorzugehen. Einen wesentlichen Teil bilden juristische Texte, die relativ leicht in größerem Maß zu beschaffen sind. Weiterhin nehmen wir für jedes Jahrzehnt zwei oder drei ‚Ratgeber' unterschiedlicher Art auf, z.B. Benimmbücher, Kochbücher und ähnliches. Vorgesehen, aber schwierig zu beschaffen sind Texte wie Theaterzettel, jedenfalls wenn man vergleichbare Texte über längere Zeiträume haben will. Was wir auch zu bekommen versuchen, sind Gebrauchsanweisungen von Firmen wie z.B. Siemens, die ja relativ lange existieren. Solche Gebrauchsanweisungen, für Staubsauger beispielsweise, können dann vergleichend aufgenommen werden. Es ist nicht ganz leicht an solche Texte zu kommen, aber wir versuchen es. Die letzte der fünf Kategorien hatte ich bereits kurz besprochen: das ist gesprochene Sprache.

Soviel zur Zusammensetzung des Kerncorpus. Ich möchte natürlich nicht behaupten, daß wir damit alles erfaßt haben, was an der deutschen Sprache des vergangenen Jahrhunderts lexikographisch interessant ist; jede solche Behauptung wäre eine Anmaßung. Es ist aber sehr viel, und es stellt eine sehr breite Abdeckung dar, wenn man es mit anderen Corpora oder auch mit den Belegsammlungen älterer Wörterbücher vergleicht.

Wie kommt man an diese Texte heran? Und wie bekommt man sie in den Computer? Da hat man zwei große Probleme. Das eine ist juristischer Art: wer gibt einem die Texte überhaupt, denn die meisten sind nicht gemeinfrei, da in Deutschland die Regel gilt, daß bis zu siebzig Jahre nach dem Tod eines Autors die Rechte nicht frei sind. Erfreulicherweise haben wir inzwischen eine sehr gute Kooperation mit ein paar einschlägigen Stellen: Wir haben beim „Spiegel" und der „Zeit" die Möglichkeit, Texte ad libitum zu bekommen. Vom ‚Spiegel' haben wird derzeit hundert vollständige Ausgaben; mehr würde einfach unser Corpus sprengen. Ähnliches gilt für die „Zeit". Beide haben uns die Rechte für die interne Wörterbuchbenutzung gegeben; die entsprechenden Texte können allerdings nicht frei ins Internet gestellt werden, so daß

andere Nutzer Zugriff darauf haben. Wir haben - was literarische Texte betrifft - eine gute Kooperation mit einer Reihe von Verlagen. Hier dachten wir uns, mit Suhrkamp anzufangen, einem Haus, das dafür berühmt ist, nicht nur über viele bedeutende Texte zu verfügen, sondern auch sehr zurückhaltend bei der Weitergabe zu sein. Und es ist uns gelungen: wir haben inzwischen die Erlaubnis und auch zum Teil die Texte von einer ganzen Reihe wissenschaftlicher, belletristischer und literarischer Autoren. Wir haben beispielsweise Werke von Habermas, Adorno, Benjamin, Bloch, aber auch z.B. Johnson, Enzensberger oder Walser - insgesamt von zweiundzwanzig Autoren des Hauses. Das Problem ist, daß Suhrkamp in der Regel die Entscheidungen nicht selbst treffen kann, sondern zuerst die Autoren anschreiben muß, da in den meisten älteren Verträgen die elektronische Verwertung gar nicht geregelt ist. Suhrkamp sieht es, wie auch andere Verlage, nicht gerne, daß wir direkt mit den Autoren Kontakt aufnehmen, sondern möchte den Zugang schon kanalisieren. Vom Aufbau-Verlag, der sehr kooperativ war, haben wir eine Reihe von Texten, z.B. Viktor Klemperers Tagebücher - ein auch in diesem Zusammenhang besonders wertvoller Text, da er sich auch über einen relativ langen und schwierigen Zeitraum erstreckt. Ferner haben wir eine grundsätzliche Zusage von Fischer, eine Reihe von Texten zu bekommen – was auch nicht leicht zu erreichen war.

Daß man diese Texte nutzen darf, heißt, wie schon bemerkt, allerdings nicht unbedingt, daß man sie sozusagen zur freien Nutzung ins Internet stellen dürfte. Es ist im Moment schwer zu sagen, wieviele Texte wir allgemein, also zur Benutzung durch jeden Interessierten, freigeben dürfen. Erlaubt ist zunächst nur die Verwendung für die Wörterbucharbeit. Im Moment verhandeln wir mit mehreren Verlagen, ob man dieses Recht nicht auf die allgemeine wissenschaftliche Nutzung ausdehnen könnte. Es ist nichts entschieden; ich bin im Moment nicht pessimistisch, aber ich glaube, viele von Ihnen kennen die allgemeine Diskussion über Urheberrechtsfragen, in der die Wissenschaft und damit die wissenschaftliche Nutzung eigentlich kaum eine Rolle spielt. Von besonderen Rechten ist da nicht die Rede, obwohl ja in der Verfassung besondere Vorrechte für die Wissenschaft vorgesehen sind, die vielleicht hier eine Rolle spielen sollten. Bevor da grundsätzliche Entscheidungen europaweit getroffen sind, wird die Situation sehr, sehr schwierig bleiben.

Das ist die eine, die rechtliche Seite. Die andere Seite ist die technische: wie bekommt man die ausgewählten Texte in den Computer? Das kommt in erster Linie darauf an, in welcher Form sie vorliegen. Wir schätzen, daß wir zwischen fünfzig und sechzig Prozent unserer Texte bereits in elektronischer Form bekommen. So liegen die Ausgaben des ‚Spiegel' und vieles andere bereits in digitalisierter Form vor. Auch die Texte von Suhrkamp, Aufbau usw. kann man großenteils in dieser Form erhalten. Die übrigen Texte müssen ‚eigendigitalisiert' werden. Das haben wir auch zum großen Teil schon getan, und zwar ungefähr in der folgenden Verteilung: ungefähr zehn Prozent der Texte werden eingescannt und dann über OCR - ‚optical character recognition' - digitalisiert. Man denkt zunächst einmal, daß dieses Verfahren das Beste ist; aber es ist extrem fehleranfällig und für ältere Texte, wie beispielsweise Frakturzeitungen von 1903, praktisch überhaupt nicht zu nutzen. Es klingt ja vielleicht beeindruckend, wenn man liest, daß 99% eines Textes richtig erkannt wird; das sind die Werte, die die Herstellung von OCR-Verfahren angeben. Aber wenn man daran denkt, daß eine normale Buchseite vielleicht 3000 Zeichen enthält, dann heißt dies, daß pro Seite 30 Fehler sind – die dann

manuell gefunden und korrigiert werden müssen. Eine Erkennungsquote von 99% ist aber nur bei sehr guten Vorlagen zu erreichen, selbst wenn es sich um Antiqua-Satz handelt. Meistens liegt die Fehlerquote wesentlich höher; das OCR-Verfahren ist also doch recht aufwendig.

Für den größten Teil der älteren Texte müssen wir daher die ‚China-connection' nutzen, d.h. wir lassen sie, wie andere auch, in China abtippen. Bei uns sind es etwa dreißig Prozent, die so bearbeitet werden. Dieses Verfahren funktioniert so, daß zwei Chinesen den Text unabhängig voneinander abtippen – übrigens ohne auch nur ein Wort zu verstehen - und dann hinterher diese zwei Versionen verglichen werden. Ich hab mir schon mal überlegt, wenn jeder Chinese, Kinder nicht gerechnet, auch nur eine Seite für uns abtippen würde, dann wären wir bei einer Milliarde Seiten, mehr als wir brauchen. Soweit sind wir leider nicht - aber immerhin sind es ungefähr dreißig Millionen Wörter laufenden Textes, die wir bis Anfang nächsten Jahres aus dem Reich der Mitte bekommen werden. Das sind, wee gesagt, durchweg ganz alte Texte, weitestgehend in Fraktur, und wir sind sehr, sehr erstaunt über die geringe Fehlerquote. Vereinbart war eine Quote von maximal drei Fehlern auf zehntausend Zeichen, also etwa einer auf eine Buchseite, und unsere Chinesen liegen besser. Das würde hierzulande kaum jemand schaffen. Man muß bedenken, daß die Schreiber nichts verstehen und daß es sich dann wohlgemerkt um Frakturtexte handelt.

Wir haben also insgesamt rund sechzig Prozent, die digital verfügbar sind, dreißig Prozent, die in China eingetippt werden, zehn Prozent, die wir einscannen und dann von Hand korrigieren. Nun genügt es nicht, die Texte in digitaler Form auf dem Computer zu haben, sondern damit man überhaupt sinnvoll lexikographisch damit arbeiten kann, müssen diese Texte aufbereitet werden. Dazu werden die Texte im XML-Format ‚annotiert' und zwar in zweierlei Hinsicht. Zum einen werden nichtlinguistische Angaben zum Text als Ganzem aufgenommen, z.B. wann er erschienen ist, wer der Autor ist und dergleichen mehr. Diese Angaben liefern Hintergrundinformationen über die einzelnen Dukumente, die nicht zuletzt auch dazu dienen können, die Texte nach verschiedenen Hinsichten in Textsorten einzuteilen und damit für eine spätere Suche sinnvoll zu filtern.

Die andere Art von Annotation ist eigentlich viel schwieriger und letzten Endes für die Wörterbucharbeit wichtiger, und das ist die linguistische. Man hat ja, wenn der Text in digitaler Form aufgenommen ist, zunächst einmal nur den reinen Wortlaut. Was man aber eigentlich bräuchte, sind lemmatisierte Texte, d.h. die verschiedenen Flexionsformen eines Wortes müssen unter einem Stichwort gesammelt werden. Außerdem hätte man gerne zumindest so etwas wie eine grobe syntaktische Analyse, so daß man etwa Nomen und Verben unterscheiden kann. Ebenso ist ein solche Analyse erforderlich, wenn man lexikalische Einheiten, die aus zwei oder mehr getrennt geschriebenen Einheiten bestehen, erfassen will. Dazu zählen als erstes alle Verben mit trennbarer Partikel, z.B. ‚vorlesen': *Er LAS mehrere Stunden aus dem Buch VOR*. Zum andern fallen alle möglichen festen Wendungen darunter, die ja einen wesentlichen Teil des lexikalischen Repertoires einer Sprache ausmachen, aber traditionell in der Wörterbucharbeit etwas zu kurz kommen. Ich komme am Schluß meines Vortrags noch einmal auf diese ‚Kollokationen' zurück.

Für die Demo-CD haben wir ursprünglich den *Stuttgarter Tagger* verwendet, der am dortigen Institut für maschinelle Sprachverarbeitung entwickelt worden ist. Ich weiß nicht, ob das Wort *Tagging* so geläufig ist: das ist eine syntaktische Analyse, die aber

eigentlich nicht tiefer geht als Wortklassenebene. Je nachdem, wie fein man diese Analyse macht, kann man bis zu fünfzig Wortklassen und Unterklassen unterscheiden - die wir nicht brauchen. Wir nehmen eine etwas grobere Analyse vor, die automatisch die Texte lemmatisiert und getrennt geschriebene Wortformen zusammen bringt. Für die Hauptphase werden wir den in Saarbrücken entwickelten Tagger M-PRO von Johannes Haller verwenden, da er für unsere Zwecke besser ist. So etwas wie einen idealen Tagger gibt es nicht. Es kommt immer sehr darauf an, welche Zwecke man damit verfolgt, und in einigen Vortests hat sich gezeigt, daß M-PRO sich für unsere speziellen Zwecke am besten eignet.

Ein ganz wichtiger Aspekt bei computerlexikographischen Unternehmen ist natürlich die Software, d.h. die Programme, mit denen man die in irgendeiner Weise aufbereiteten Texte für den Benutzer erschließbar macht. Das große Problem liegt hier nicht allein in der Entwicklung geeigneter Verfahren, für die man versierte Programmierer braucht, sondern vor allem darin, die Aktualität der Software über einen längeren Zeitraum zu gewährleisten. Sie wissen alle, daß Programme, die vor zehn Jahren entwickelt wurden, in vielen Fällen für den eigenen PC schon überhaupt nicht laufen. Das wäre für ein langfristig angelegtes Wörterbuchprojekt fatal. Wir haben uns von Anfang an entschlossen, möglichst wenig eigenständige Programme zu schreiben und einzusetzen, sondern an Softwareentwicklungen anzuschliessen, von denen man erwarten kann, daß sie über einen längeren Zeitraum gepflegt und von sehr vielen Leuten in verschiedenen Institutionen benutzt werden. Die erste Idee, die wir für unsere Demo-CD hatten, war es, die aus dem Internet bekannte Suchmaschine ALTAVISTA zu verwenden. Nun liefert ALTAVISTA, wie eigentlich all die großen Suchmaschinen, nur die Angabe, daß ein bestimmtes Wort, genauer gesagt, eine bestimmte flektierte Wortform, in einem Dokument auftaucht. Es findet aber nicht die Stelle und damit den genauen Beleg samt Kontext. Damit kann man keine Konkordanzen und dergleichen machen, ebensowenig kann man lemmatisieren oder ‚taggen'. Wir haben also noch ein eigenes Programm draufgesetzt, das in der Tat dies leistet. Dieses Programm ist eine Eigenentwicklung, aber eine vergleichsweise einfache, die sich auch leicht durch andere Programme mit ähnlicher Leistung ersetzen läßt. Sie können sich selbst ein Bild davon machen, wenn Sie sich einmal an die genannte Internetadresse www.dwds.de wagen und in dem Democorpus herumstöbern.

Wir haben uns jetzt aus einer ganzen Reihe von Gründen aber entschlossen, stattdessen die neueste Version der bekannten Datenbank-Software ORACLE als Grundlage zu nehmen. Sie ist extrem leistungsfähig, kann riesige Textmengen verwalten – das *Spiegel*-Archiv beispielsweise, das von einer ORACLE-Datenbank verwaltet wird, umfaßt meines Wissens derzeit ungefähr 30 Milliarden Textwörter -, und ist vor allem weltweit bei sehr vielen Institutionen im Einsatz; daher kann man einigermaßen sicher sein kann, daß ORACLE auch noch in zehn Jahren zur Verfügung stehen wird, und wenn es hin und wieder zu einer Umstellung kommt, dann wird sie sehr viele sehr viel größere Einrichtungen betreffen; dann können wir uns anschließen. Es wird speziell für unsere Zwecke noch ein sogenannter ‚Lexer' hinzugefügt – eine Verbindung von Lemmatisierer und Tagger, der voraussichtlich Anfang nächsten Jahres fertig sein wird. Wir haben auch festgestellt, daß diese Software außerordentlich schnell ist. Ein großes Problem bei großen Datenbanken ist natürlich, daß sie langsam werden. Man denkt sich, daß es im Prinzip keinen Unterschied macht, eine Datenbank mit zehn Millionen, mit hundert

Millionen oder mit tausend Millionen Wörtern zu verwalten und zu durchsuchen. Aber es ist ungefähr so, als ob man aus 3 Metern, aus 30 Metern oder aus 300 Metern Höhe herabspringt. Wir haben unsere Version von ORACLE mit dem ‚opportunistischen Corpus' – derzeit eine halbe Milliarde Wörter - ausprobiert, und es ergeben sich wirklich nutzbare Zeiten von Millisekunden, allenfalls Sekunden, um irgend ein Wort oder eine Wortverbindung zu suchen.
Soweit zu Zusammensetzung, Aufbereitung und Erschließung des Kerncorpus. Die Arbeiten daran sollen Mitte nächsten Jahres fertig sein. Ich glaube nicht, daß wir zu 100% schaffen, was wir uns vorgenommen haben, vor allem weil die Textbeschaffung manchmal doch ein wenig schwierig ist. Aber im Großen und Ganzen wird das Kerncorpus, das ja die Grundlage des DWDS ist, Mitte 2002 stehen und für lexikographische Zwecke nutzbar sein.
Ich komme jetzt zur zweiten Phase, nämlich der lexikographischen Ausarbeitung und damit zur eigentlichen Arbeit des Lexikographen. Ein umfangreiches, gut ausgewähltes Corpus ist auch für sich genommen eine schöne Sache, vor allem wenn es sorgfältig aufbereitet ist und man effizient darin suchen kann. Ich denke auch, daß nicht nur Wissenschaftler, sondern beispielsweise Journalisten oder Übersetzer, die nach irgendeiner Verwendung eines Ausdrucks suchen, einen großen Nutzen aus einem derartigen Textbestand ziehen können. Aber eigentlich möchte man mehr haben: man möchte eine lexikologische oder lexikographische Analyse haben. Wie kann man das machen? Eines ist zunächst einmal klar: es ist nicht zuletzt ein Finanzproblem. Ein klassisches Langzeitvorhaben ist ausgeschlossen, denn keine Akademie – oder jedenfalls keine deutsche Akademie - wird ein Langzeitvorhaben in dieser Form noch einmal bewilligt bekommen. Das also geht nicht. Was kann man da tun? Die Idee, die wir hatten, ist, auch hier die Möglichkeiten des Computers voll zu nutzen. Anders als beim traditionellen Vorgehen ist es hier nicht nötig, das Wörterbuch von A bis Z auszuarbeiten, sondern man kann schrittweise einzelne, in sich abgeschlossene Komponenten entwickeln, die dann auch schrittweise finanziert werden können. Der Abschluß einer solchen Komponente (wir nennen sie übrigens modischerweise ein ‚Modul' und reden von einem ‚modularem Aufbau') hat dann schon zum Ergebnis, daß das gesamte Datenmaterial in einem bestimmten Bereich sinnvoll nutzbar gemacht wird, während andere Bereiche dann später bearbeitet werden.

Wie hat man sich das im einzelnem vorzustellen? Ich denke, daß dies vielleicht der rechte Zeitpunkt ist, sich einmal ganz grundsätzlich zu überlegen, was es eigentlich heißt, den Wortschatz einer Sprache lexikographisch zu erschließen. Es ist mir selbst schon mehrfach aufgefallen, wie sehr meine eigene Denkweise über die Beschaffenheit des Wortschatzes einer Sprache von der Vorstellung des gedruckten Wörterbuches geprägt worden ist – von der Vorstellung einer alphabetisch oder manchmal auch nach anderen Kriterien geordneten Liste. Aber das ist halt nur ein durch das Medium, den Buchdruck, vorgegebener Weg, das, worum es eigentlich geht, nämlich den Wortschatz der Sprache, linear anzuordnen. Der Wortschatz selbst ist natürlich nicht so strukturiert. Es scheint mir wichtig, sich dies einmal ganz grundsätzlich vor Augen zu führen. Seit den Tagen der griechischen Grammatiker unterscheidet man eigentlich immer in der Beschreibung einer Sprache zwischen ‚Lexikon' und ‚Grammatik', wobei diese Ausdrücke mehrdeutig sind zwischen dem, was beschrieben wird, und der Beschreibung selbst. Es gibt so etwas wie elementare Ausdrücke, die die ‚Lexik' oder das ‚Lexikon' der Sprache ausmachen und

demnach in einem Verzeichnis, eben dem ‚Wörterbuch' oder dem ‚Lexikon' stehen; und es gibt eine Menge von Regeln, die festlegen, wie man aus bestehenden Ausdrücken neue, morphologisch oder syntaktisch komplexere Ausdrücke bilden kann; diese Regeln bilden die ‚Grammatik', und sie werden in der ‚Grammatik' beschrieben. Nun brauche ich Ihnen nicht zu sagen, daß der Übergang zwischen Lexikon und Grammatik gleitend ist: ein Beispiel hierfür sind feste Wendungen, d.h. Ausdrücke, die der Bedeutung nach eine Einheit bilden, aber der Form nach nach irgendwelchen syntaktischen und morphologischen Regeln gebildet werden. Dennoch, grundsätzlich kann man diese Unterscheidungen so machen. Es geht also darum, die Lexik einer Sprache zu beschreiben.

Was ist eigentlich die Lexik einer Sprache? Es ist im Grunde eine Ansammlung von sprachlichen Einheiten, die durch bestimmte Relationen miteinander verbunden sind. Eine solche Einheit – ich sage der Einfachheit halber ‚Wort', obwohl es nicht immer Wörter im Sinne von ‚zusammengeschriebene Einheiten' sind -, eine solche Einheit also ist letzten Endes nichts anderes als eine Verbindung von verschiedenen Formen von Information. Ein Wort ist ja nicht das, was man hinschreibt, sondern eine Wort ist ein Bündelung von phonologischen Informationen, graphematischen Informationen, von morphologischen Informationen, syntaktischen Informationen und semantischen Informationen. Ich weiß, daß dies für Sie als Sprachwissenschaftler etwas trivial ist, möchte aber trotzdem, daß sie sich es genau so noch einmal vor Augen führen, denn manchmal vergißt man das Offenkundige. Die Idee, die wir haben, ist nun, nicht nach einem alphabetischen oder ähnlichen Linearisierungsprinzip vorzugehen, sondern jeweils diese Arten von Informationen, auf Grund deren sich auch die Beziehungen der Einheiten in der Lexik definieren lassen, als Ausgangspunkt der einzelnen Arbeitsschritte zu nehmen. Wie kann man beispielsweise die phonologische Information des deutschen Wortschatzes der Gegenwart am besten erfassen? Normalerweise geschieht dies über eine Lautschrift, heute meistens die Lautschrift der API. Der Plan, den wir für dieses Modul verfolgen, ist es nun, nicht eine solche Umschrift zu verwenden, sondern einfach ein ‚Soundlinking' bei den einzelnen Wörtern zu machen. Das heißt also, daß man sich das Wort einfach auf einen Klick hin anhören kann, und zwar in verschiedenen Varianten – also z.B. in wienerischer Aussprache, in Züricher Aussprache oder von einem Tagesschausprecher gelesen. Die Idee ist also, die phonologische Informationen nicht zu ‚graphematisieren', sondern sie über den Schall selbst zugänglich zu machen. Eigentlich ist das ja auch viel natürlicher. Es ist in gewisser Weise so, als würde man in einem gedruckten Wörterbuch mit dem Finger über ein Wort streichen, und schon hört man es. Mit einem klassischen Wörterbuch ist das natürlich nicht zu machen, aber mit den Möglichkeiten des Computers ist es kein großes Problem. Am Max-Planck-Institut für Psycholinguistik, an dem ich arbeite, haben wir sehr viel Erfahrung mit diesem Soundlinking von Corpora in verschiedenen Sprachen gesammelt.

Die morphologische und die syntaktische Information, zwei weitere Module also, kann man relativ leicht durch Ausnutzung von inzwischen verfügbaren syntaktischen und morphologischen Analyseverfahren erschließen. Die Morphologie ist sowieso kein großes Problem, jedenfalls soweit es um die Flexion geht; letztlich geht es ja nur darum, die Flexionsklassen anzugeben. Eine der schönen Möglichkeiten des Computers ist allerdings, daß man auch wiederum auf einen Klick hin alle Flexionsformen eines Eintrags auch *erzeugen* lassen kann, und dies in allen belegten Varianten. Im Prinzip

kann man sich auch gleich die Verteilung dieser Formen über die Zeit oder über die Textsorten angeben lassen. Ein bißchen schwieriger ist es mit der Syntax. Es gibt eine Reihe von automatischen Analyseverfahren, sogenannten Parsern, von denen aber keiner auch nur annähernd fehlerfrei funktioniert, sobald es um eine gewisse oberflächliche Analyse hinausgeht. Man muß sich vor der Vorstellung hüten, man könne sozusagen das ganze Corpus durch einen Parser jagen, und dann hätte man es hinterher so gut syntaktisch analysiert, daß man daraus die relevanten syntaktischen Informationen eines Wortes ableiten kann. Immerhin kann sich die Arbeit durch eine solche automatische Analyse ganz massiv erleichtern; es ist nur sehr viel manuelles Nacharbeiten erforderlich. Generell stellt sich dabei natürlich die Frage, wieviel syntaktische Information man in ein Wörterbuch stecken will; ein Wörterbuch ist schließlich keine Grammatik. Aber viele grammatische Regeln sind ‚wortbasiert', d.h. an die lexikalischen Eigenschaften bestimmter Wörter gebunden.

Das schwierigste in der lexikalischen und lexikographischen Analyse ist natürlich die Bestimmung der semantischen Informationen, die zu einem Wort gehören. Auch hier kann man sich die Arbeit in gewissen Grenzen mit Mitteln des Computers leichter machen, weil es viel einfacher ist, die Belege durch die Zeiten und durch die Textsorten zu verfolgen als über Zettelkästen. Man kann sich beliebig viel Kontext geben lassen, kann alle Verwendungen eines Wortes bei einem Autor in einem bestimmten Werk nebeneinanderstellen usw. All dies erleichtert die Arbeit sehr. Aber letztlich führt kein Weg daran vorbei, daß die Bedeutung oder die Bedeutungen einer lexikalischen Einheit ‚per Kopf' bestimmt werden müssen. Das geht nur über die linguistische Kompetenz des Lexikographen, und das braucht Zeit. Wir haben, was das ‚semantische Modul', also das Herzstück der lexikographischen Arbeit angeht, noch keine festen Vorstellungen, allerdings wohl einige Vorüberlegungen. Insbesondere denken wir, daß man auch dies auf keinen Fall ‚von A bis Z' machen sollte. Vielmehr wir haben die Vorstellung, daß man den Wortschatz in einzelne Bereiche zerlegt, die getrennt bearbeitet werden, semantische Teilmodule sozusagen – beispielsweise die Bewegungsverben, das System der deutschen Ortspräpositionen, oder, wenn dies finanzierbar ist, zunächst einmal alle ‚Grundwörter' (d.h. alle morphologisch einfachen Wörter), eventuell erst von einer bestimmten Häufigkeit an. Man kann sich durchaus denken, daß bestimmte Teile in Form von Magisterarbeiten oder Dissertationen ausgearbeitet werden. Auf diese Weise wird nach und nach eine Komponente nach der anderen hinzufügt, deren jede ist für bestimmte Zwecke schon benutzbar ist - und zwar sehr schnell benutzbar! Vor allem aber ist es so möglich, Finanzierungsträger zu finden, die bereit sind, dies zu zahlen. Niemand ist heute bereit, ein ‚Jahrhundertprojekt' zu finanzieren – aber es ist durchaus möglich, das Geld für einzelne, in sich abgeschlossene Komponenten zu finden. Dieses Vorgehen hat auch den Vorzug, daß man jederzeit immer wieder Informationen hinzufügen kann, neue Belege, Korrekturen und dergleichen. Es ist einfach sehr flexibel und sehr dynamisch. Aber man wird eben natürlich nicht das ganze Wörterbuch von A bis Z haben.

Lassen sie mich jetzt zu einer letzten Komponente kommen, die ich bislang nicht erwähnt habe, weil sie ein wenig quer zu den übrigen liegt; dies sind ‚feste Wendungen', also Ausdrücke, die syntaktisch komplex sind, sich der Bedeutung nach aber wie ein einzelnes Wort verhalten. Sie bilden, wie Sie alle wissen, eines der größten Probleme der lexikographischen Arbeit, und trotz einiger bemerkenswerter Leistungen auf diesem Gebiet ist es sicher fair zu sagen, daß sie nach wie vor ein Stiefkind sind – es gibt keine

wirklich befriedigende Aufarbeitung der ‚Kollokationen' der deutschen Sprache. Kollokationen ist hier – Sie wissen, es gibt auf diesem Gebiet viele terminologische Diskussionen - einfach als Oberbegriff für die verschiedenen Formen fester Wendungen zu verstehen. Wir wissen auch alle, daß das ein gleitender Bereich ist; das brauche ich in diesem Kreis gar nicht näher zu erläutern. Eines unserer Ziele war es, die Kollokationen als ein eigenes ‚Modul' zu behandeln, das für sich bearbeitet und für sich finanziert werden soll. Wie einige in der letzten Zeit in der Zeitung gelesen haben werden, ist es uns in der Tat gelungen, für diesen Bereich eine Finanzierung zu bekommen, und zwar in Form einer Kooperation mit Christiane Fellbaum von der Princeton University, die dort das berühmte ‚Wordnet-System' mitentwickelt hat. Ihr wurde jetzt von der Alexander-von-Humboldt-Stiftung der ‚Wolfgang-Paul-Preis' verliehen, und zwar für eine Analyse der Kollokationen der deutschen Gegenwartssprache gemeinsam mit der BBAW. Nicht alle Details sind bereits festgelegt, aber es gibt inzwischen eine Reihe von Konkretisierungen. Anfangen werden wir mit Verb-Nomen-Kollokationen, und zwar unabhängig davon, ob das Nomen als Objekt oder als Präpositionalphrase oder dergleichen realisiert ist, also Ausdrücke wie: *auf die Nerven fallen, die Hufe hochreißen* oder *Trübsal blasen* und dergleichen mehr. Das ist natürlich nur ein Teilbereich der Kollokationen im Deutschen, aber ein wichtiger, und er soll so umfassend wie möglich abgedeckt werden. Das heißt, daß in diesem Bereich der Verb-Nomen–Verbindungen *alle* Kollokationen untersucht werden sollen, die wir in unserem Kerncorpus haben. Das ist sehr viel – hundert Millionen Wörter sind im Prinzip hundert Millionen Belege, und selbst wenn man die häufigen Funktionswörter abrechnet, ist es sicherlich das zehnfache dessen, was in den großen historischen Wörterbüchern wie DWB oder OED erfaßt wurde. Die schiere Masse ist natürlich nur ein Problem.

Ein zweites und schwierigeres ist, daß es eben, wie gerade bemerkt, keine klare Abgrenzung zwischen festen Wendungen und frei komponierten, also ganz produktiven Wendungen gibt. Wir werden deshalb also ganz vorsichtig, gleichsam experimentell, anfangen und in der ersten Phase, die sehr schnell beginnen soll, von den Verben ausgehen; wir werden für die zweihundert häufigsten deutschen Verben (Wörter wie *gehen* und *stellen*) sämtliche Verbindungen untersuchen, die wir in unserem Corpus finden und die ‚kollokationsverdächtig' sind, und diese dann nach syntaktischen und semantischen Gesichtspunkten beschreiben. Auf Grund der Erfahrungen, die wir damit gemacht haben werden, wollen wir dann sehen, inwieweit wir dieses Vorgehen auf andere Verbindungen übertragen können. Die Gerechtigkeit gebietet zu sagen, daß man nicht vorhersagen kann, wie gut das im Einzelnen laufen wird. Wir werden es auf jeden Fall so versuchen, und die ersten Erfahrungen damit, erst einmal ein bißchen rumzuprobieren, stimmen mich eigentlich sehr optimistisch. Das Projekt soll im kommenden Jahr beginnen und im Laufe von drei Jahren abgeschlossen werden. Damit hoffen wir also, ein ‚Modul' - die Bearbeitung der Kollokation in diesem umschriebenen Bereich – fertigstellen zu können. Wir werden im Moment keine anderen Komponenten in Angriff nehmen. Phonologie beispielsweise wäre sehr interessant und könnte gemacht werden, aber ich glaube, es würde einfach unsere Arbeitskraft übersteigen.

Damit habe ich Ihnen zumindest eine gewisse Vorstellung davon vermittelt, was das ‚Digitale Wörterbuch der deutschen Sprache des 20. Jahrhunderts' ist oder, besser gesagt, was es sein soll. Was die BBAW hier in Angriff genommen hat, ist eigentlich kein Wörterbuch im üblichen Sinne. Es gibt kein gutes Wort dafür. Ein lexikalisches System?

Ein *Lexon* könnte man vielleicht sagen. Ich weiß es nicht. Vielleicht sollte man ein Preisausschreiben für die beste neue Bezeichnung ausloben. Wie immer man es nennt - es ist ein System, das den Wortschatz einer bestimmten Sprache in bisher unerhörter Dichte und Flexibilität zu beschreiben erlaubt. Mir scheint, es ist schon eine gewaltige Veränderung in der ganzen Tradition der Lexikographie und der Lexikologie seit Calepinus, die sich hier nicht nur bei uns, sondern auch bei manchen anderen Institutionen abzeichnet.

Lassen sie mich schließen mit einer Bemerkung, die ich neulich noch einmal mit großem Vergnügen gelesen habe. Ich denke, vieles hat sich geändert in der Wörterbucharbeit, aber eines hat sich nicht geändert, das nämlich, was Samuel Johnson vor zweihundertfünfzig Jahren im Vorwort zu seinem berühmten Wörterbuch geschrieben hat. Nämlich, daß man in allen möglichen Disziplinen als Wissenschaftler nach Ruhm und Ehre streben darf, daß aber ein Lexikograph allenfalls hoffen kann, dem Tadel zu entgehen: ‘Every other author may aspire to praise; the lexicographer can only hope to excape reproach.’