Klein, W. (2004). Vom Wörterbuch zum Digitalen Lexikalischen System. *Zeitschrift für Literaturwissenschaft und Linguistik, 136,* 10-55.

Aus Zeitschrift für Literaturwissenschaft und Linguistik 136, 2004, 10-55

Wolfgang Klein

Vom Wörterbuch zum Digitalen Lexikalischen System

> ... welch lange und mühselige Arbeit hat dieses Werk mir auferlegt, welchem Gram und Kummer, welchen Kränkungen und Verletzungen mich ausgesetzt, welche Opfer von mir gefordert! Gesundheit, Besitz und Erwerb habe ich für dasselbe hingeben müssen; ja selbst der Fürsorge für die Meinigen hat es mich beraubt, indem es mich auf jeden Nebenverdienst, durch den ich, wenn auch nicht die Zukunft meiner Familie sicherstellen, doch ihr Schicksal erleichtern konnte, Verzicht zu leisten verpflichtet hat. Nur durch frommes, vertrauensvolles Gebet und durch treuen unermüdlichen Fleiß - zu ermutigendem Trost sei dies allen gesagt, denen es, gleich mir, auf ihrem Weg zu einem fernen Ziele an Hilfe gebricht - bin ich, wenn auch spät, erst beim Sinken meines Leben, halberblindet und an Geist und Körper geschwächt, der Vollendung meines Werkes nahegekommen.
>
> Eberhard Gottlieb Graff, Vorwort zum "Althochdeutscher Sprachschatz" (Berlin 1834)

1. Einleitung

War es das wert? Eberhard Gottlieb Graff, zu dieser Zeit 54 Jahre alt, ist sieben Jahre später, ein Jahr vor Veröffentlichung des sechsten und letzten Bandes, gestorben. Sein monumentales Wörterbuch ist mit Recht gerühmt worden; aber es ist, weil nach Wurzeln statt alphabetisch geordnet, ohne weitere Hilfsmittel - Maßmann, eben jener Maßmann, den Heine so gnadenlos verspottete, hat einige Jahre später einen alphabetischen Index nachgereicht - fast nicht zu benutzen. Es ist akademisch in den verschiedensten Schattierungen des Wortes: gründlich, kenntnisreich, von einer Sorgfalt, die wenig pragmatische Kompromisse kennt, gedacht für wenige, von diesen wenigen kaum genutzt. Es bezeichnet den einen Pol in der mehr als viertausendjährigen Geschichte der Lexikographie, an deren anderem Pol der Polyglottsprachführer steht, der den Türkeireisenden vor den übelsten Fährnissen des Alltags in Antalya bewahren soll.

So wie die Mathematik, die reinste aller Wissenschaften, ihren Ausgang aus der Feldvermessungen und den Bedürfnissen des Handels genommen hat, so stehen am Anfang der Lexikographie praktische Zwecke: Tontäfelchen, in denen zweieinhalbtausend Jahre vor Christi Geburt sumerisch-akkadische Schreiber tausende von sumerischen Wörtern mit ihren akkadischen Gegenstücken verzeichnet haben. Das Land war zweisprachig, für den Handel und den Wandel war es hilfreich, die Bezeichnungen in der jeweils anderen Sprache zu kennen, und so hat man sie denn aufgeschrieben, sobald man überhaupt schreiben konnte. Seither hat jede Schriftkultur Wörterbücher hervorgebracht. Die Lexikographie ist nicht nur die älteste, sondern mit weitem Abstand die wichtigste sprachwissenschaftliche Disziplin. Ihre Aufgabe ist es immer, einen Ausschnitt des Wortschatzes einer Sprache darzustellen. Die konkreten Umsetzungen dieses Ziels - die Wörterbücher - könnten nach Form, Umfang, Zweck und Brauchbarkeit, um nur einige der Unterscheidungsmerkmale zu nennen, nicht unterschiedlicher sein. Eines aber ist ihnen immer gemeinsam: es sind Bücher oder, etwas allgemeiner gesagt, buchähnliche Gegenstände. Die Darstellung des Wortschatzes ist eine Folge zweidimensionaler Seiten, ganz gleich, ob diese gebunden, aneinandergeklebt und gerollt oder einfach auf- oder nebeneinandergelegt sind wie beim Tontäfelchen. Buchrolle, Buch und Tontäfelchensammlung sind als als physikalische Objekte natürlich dreidimensional; aber das ist für die Darstellung des Wortes irrelevant; dafür zählt nur "Oben-Unten, Links-Rechts"; ich werde deshalb weiter unten von einer OULR-Darstellung reden. Entscheidend ist aber immer, daß man zur Repräsentation all dessen, was man über ein Wort sagen will, auf zwei Dimensionen, zumeist eben eine Buchseite, angewiesen ist. Dieser Zwang des Formats bestimmt die Darstellung des Gegenstandes, ja, sie prägt sogar in hohem Maße sogar die Vorstellung davon, was denn dieser Gegenstand eigentlich ist: ein Wort ist etwas Geschriebenes, mit dem sich weitere Informationen verknüpfen lassen. Aber das Geschriebene ist natürlich nur ein Aspekt des Wortes, und nicht einmal ein konstitutiver: die meisten Sprachen in der Geschichte der Menschheit sind schriftlos, und sie haben auch Wörter. Es

wäre auch schwer, dem Zwang der Zweidimensionalität zu entgehen, solange es keine anderen wirklich brauchbaren Möglichkeiten der Darstellung gibt.

Dies hat sich mit dem Aufkommen des Computers geändert. Die Repräsentation irgendwelcher Informationen im Computer braucht nicht zweidimensional zu sein. So eröffnen sich dem Lexikographen prinzipiell neue Möglichkeiten, Ausschnitte aus dem Wortschatz einer Sprache, oder auch mehrerer Sprachen, darzustellen. Von diesen Möglichkeiten haben wir bislang nur eine ungefähre Ahnung. Zwar gibt es seit gut drei Jahrzehnten zahlreiche Versuche, den Computer für die "Wörterbucharbeit" - wie charakteristischer Weise gesagt wird -, nutzbar zu machen, insbesondere für die Corpuserstellung. Aber alles, was darüber hinausgeht, bleibt stark der Vorstellung des klassischen "Wörterbuchs" verhaftet. Es sind dies weitgehend in ein anderes Medium übertragene Wörterbücher, exemplifiziert beispielsweise an den kleinen, tragbaren Reiseübersetzern wie auch an den digitalen Versionen großer Wörterbücher wie des *Oxford Dictionary of English*(http://dictionary.oed.com), des noch von den Brüdern Grimm begonnenen *Deutschen Wörterbuchs* (http://www.dwb.uni_trier.de/index.html) oder des *Trésor de la Langue Française* (http://atilf.atilf.fr/tlf.htm). Was erheblich verbessert wurde, sind vor allem die Suchmöglichkeiten. Darüber hinaus gibt es in letzten Jahren nun eine Reihe von Versuchen, die Beschränkungen der tradierten Darstellungsform lexikographischer Information zu überwinden. Der auffälligste darunter die "Sprachausgabe" ist, d.h. die Möglichkeit, sich die Aussprache eines Wortes anzuhören statt sie mühselig aus einer Lautschrift zu erschließen. Es ist dies aber keineswegs die einzige Möglichkeit (einen guten Überblick geben die Beiträge in Lemberg u.a. 2001, insbesondere Storrer 2001).

Das "Wörterbuch" der Zukunft kann sich von den tradierten Zwängen frei machen: es ist kein Buch mehr, sondern ein flexibles, jederzeit erweiterbares und auf bestimmte Zwecke zuschneidbares *Digitales Lexikalisches System* (DLS), das in seiner Grundform nur auf dem Computer steht. Zu bestimmten praktischen Zwecken kann man daraus gedruckte Wörterbücher ableiten, die dann den tradierten Beschränkungen unterliegen, aber vielleicht andere Vorteile haben. Dieser Aufsatz ist eine Reise in das Land dieser Möglichkeiten. Dabei stütze ich mich vielfältig auf Überlegungen und Vorarbeiten, die in dem Projekt *Digitales Wörterbuch der Deutschen Sprache des 20. Jahrhunderts* (DWDS), die seit einigen Jahren an der Berlin-Brandenburgischen Akademie der Wissenschaften angestellt und durchgeführt worden sind (www.dwds.de).[1] Diese Vorarbeiten bilden die Grundlage zu einem Digitalen Lexikalischen System, das in den kommenden Jahren entwickelt werden soll.

Um den Übergang vom Wörterbuch zum Digtalen Lexikalischen System zu verstehen, muß man sich zuerst einmal vergegenwärtigen, was die Aufgaben des Lexikographen sind und welchen Beschränkungen durch das Format sie unterliegen; damit befassen sich die Abschnitte 2 - 4. Anschließend gehe ich in ganz unterschiedlicher Tiefe auf die neuen Möglichkeiten ein (Abschnitte 5 - 11). Manche der Überlegungen in diesen Abschnitten sind bereits recht konkret, andere noch ganz spekulativ.

2. Wozu Wörterbücher?

When I feel inclined to read poetry, I take down my dictionary.

Es gibt viele Gründe, den Wortschatz einer Sprache ausschnittsweise beschreiben zu wollen; aber wenn man sich die Geschichte der einschlägigen Bemühungen anschaut, so scheinen vor allem drei maßgeblich gewesen zu sein:

2.1 Praktische Bedürfnisse

Solcher Bedürfnisse gibt es viele. Aber die meisten kann man auf drei Hauptmotive zurückführen, die einander nicht ausschließen, oft gemeinsam wirken und meist nicht klar zu trennen sind:

A. Der Einzelne hat mit mehr als einer Sprache zu tun

Die schon erwähnten akkadisch-sumerischen Tontäfelchen, die überhaupt am Anfang der Lexikographie stehen, wurden schon erwähnt. Die konkreten kommunikativen Probleme, die sich aus der Mehrsprachigkeit ergeben, sind bis heute für jeden normalen Menschen das Hauptmotiv der Lexikographie. Dies gilt für den, der eine andere Sprache lernen will, für den, der sie zu eng umrissenen Zwecken braucht, wie etwa den Reisenden oder den Wissenschaftler, es gilt für den, der in der Regel beide Sprache schon sehr gut beherrscht und mit einem besonderen Problem konfrontiert ist, wie den Übersetzer. Letzeres führt uns auf den zweiten Grund.

B. Manche Wörter sind auch für den, der die Sprache kennt, schwierig

[1]Ich danke Manfred Bierwisch, Christiane Fellbaum, Alexander Geyken, Gerald Neumann, Angelika Storrer, Hartmut Schmidt und Ralf Wolz für viele hilfreiche Diskussionen im Rahmen dieses Vorhabens.

Am Anfang vieler Wörterbuchtraditionen stehen Glossen und Glossare: an den Rand eines Textes wird eine kleine Erklärung geschrieben, die ein schwieriges Wort oder auch eine schwierige Sache erläutert; diese Glossen werden alphabetisch, thematisch oder auch gar nicht geordnet zu einem Glossar zusammengestellt. Im Falle des ältesten deutschen Textes, der uns überliefert ist, dem *Abrogans,* gelten die Erläuterungen lateinischen Wörtern - es ist eine Vorstufe eines zweisprachigen Wörterbuchs. Es können aber auch Wörter in der eigenen Sprache problematisch sein. Der Titel von Robert Cawdreys englischem Wörterbuch von 1604, das allgemein als erstes einsprachiges Wörterbuch des Englischen angesehen wird, bringt es schön auf den Punkt (vgl. zur Tradition der älteren englischen Wörterbücher Hüllen 1999, der allerdings Cawdrey nicht behandelt):

> A Table Alphabeticall, conteyning and teaching the true writing, and understanding, of hard, usuall English wordes, borrowed from the Greeke, Hebrew, Latin, or French, &c.
> With the interpetation thereof by plaine English words, gathered for the benefit & helpe of Ladies, Gentlewoman, or any other unskilfull persons.
> Whereby they may the more easilie and better understand many hard English wordes, which they shall heare or read in Scriptures, Sermons, or elsewhere, and also be made able to use the same aptly themselves.

Cawdrey und seinen pragmatischen Landsleuten ist es noch hundert Jahre lang nicht in den Sinn gekommen, Wörter wie *hand* oder *of* in ein Wörterbuch aufzunehmen. Sie sind nicht schwierig, oder werden jedenfalls nicht dafür gehalten.

C. Man muß dem Verfall der Sprache widerstreiten

Am Anfang der indischen Wörterbuchtradition, nicht so alt wie die mesopotamische, aber älter als die abendländische, stehen religiöse Gründe. Die heiligen Texte sind in Sanskrit geschrieben, aber ein halbes Jahrtausend später hatte sich die Sprache so verändert, daß das Verständnis dieser Texte nicht mehr gesichert war; deshalb mußte für religiöse Zwecke der alte Stand festgehalten und festgefroren werden. Man weiß, daß Zaubersprüche und Gebete nicht wirken, wenn sie nicht richtig gesprochen oder geschrieben werden. Solche religiös motivierte Kodifizierungen einer Sprache gegen ihre natürliche Entwicklung finden sich in vielen Kulturen. Wie so viele andere religiöse Vorstellungen haben sie in säkularisierter Form überlebt, nämlich der Überzeugung, daß die Sprache verfällt und davor bewahrt werden muß. Im 1747 veröffentlichten "Plan of a Dictionary of the English Language" zu Samuel Johnsons Wörterbuch, einem Plan, der ungleich vielen anderen auch verwirklicht wurde und dessen Umsetzung eine der beeindruckensten lexikographischen Leistungen aller Zeiten ist (Johnson 1755), heißt es:

> This ... is my idea of an English Dictionary, a dictionary by which the pronounciation of our language may be fixed, and its attainment facilitated; by which its purity may be preserved, its use ascertained, and its duration lenghtened. (zitiert nach Jackson 2002, 44; siehe auch Horgan 1994)

Johnson selbst haben zunehmend Zweifel befallen, ob dies ein sinnvolles Ziel ist; im *Preface* von 1755 sagt er dazu:

> (I) now begin to fear that I have indulged expectation that neither reason nor experience can justify. When we see men grow old and die at a certain time one after another, from century to century, we laugh at the elixir that promises to prolong life to a thousand years; and with equal justice may the lexicographer be derided, who being able to produce no example of a nation that has preserved their words and phrases from mutability; shall image that his dictionary can embalm his language, and secure it from corruption and decay, that it is in his power to change sublunary, and clear the world at once from folly, vanity, and affection. (zit. nach Jackson 2002, 45).

Immerhin - der große Gelehrte bezweifelt nicht grundsätzlich, daß es wünschenswert ist, die Sprache in ihrer bestehenden Form zu bewahren (kurz darauf folgt die berühmte Bemerkung "tongues, like governments, have a natural tendancy to decay"). Unter seinen Nachfolgern hat die "Bewahrung der Reinheit" als Motiv der Wörterbucharbeit zunehmend an Gewicht verloren - nicht so unter den Benutzern. Der krasseste Beleg ist die öffentliche Aufnahme von "Webster's Third", der Neubearbeitung des bedeutendsten amerikanischen Wörterbuch überhaupt, die im Jahre 1961 erschienenen ist: von allen Kennern als überragende Leistung gelobt, wurde sie von Journalisten und Schriftstellern von Grund auf verrissen, weil es die althergebrachte Reinheit der Sprache dem Pöbel zum Opfer bringt (Morton 1994).

Die Bewahrung der Reinheit ist eines der großen normativen Momente, die traditionell in die Wörterbucharbeit eingehen. Das andere ist die Standardisierung, die nicht so sehr den praktischen Bedürfnissen des Einzelnen dient, sondern dem Zusammenhalt einer sozialen Gruppe.

2.2 Standardisierung

In der Welt gibt es derzeit rund 5000 Sprachen, darunter viele in vielen Varianten - Was ist schon Deutsch? - , und etwa 200 Nationen. Im Schnitt entfallen also 25 Sprachen auf eine Nation - bei erheblicher Schwankung im einzelnen. Es liegt im Interesse der Herrschenden, und etwas weniger auch der Beherrschten, wenn es eine "Standardsprache" gibt, die für Zwecke der Verwaltung, des Handels und dergleichen mehr von allen, von vielen oder auch nur von einer bestimmten Kaste beherrscht wird. Der wichtigste Schritt dazu ist ein Wörterbuch. In der vierten der großen alten Hochkulturen - neben der mesopotamischen, der griechisch-lateinischen und der indischen - war dies das Hauptmotiv zur Entstehung der Lexikographie. Auf Wörterbüchern wie dem im ersten vorchristlichen Jahrhundert in China entstandenen *Shuowen* beruhte buchstäblich über Jahrhunderte hinweg die Ausbildung der Gelehrten und damit das unerhört stabile Herrschaftssytem des Reichs der Mitte. Nie kam der Lexikographie größere gesellschaftliche Bedeutung zu.

Es ist dies wohl der wichtigste, aber doch nur einer von vielen vergleichbaren Fällen. Als nach dem letzten großen Aufflackern der antiken Tradition zu Beginn der Neuzeit das Latein zunehmend den Volkssprachen Platz machen mußte, ergab sich auch die Notwendigkeit, diese Volkssprachen für viele der Zwecke, denen zuvor das Latein gedient hatte, geeignet zu machen. Hier ging es nicht um die Bewahrung einer Sprache, sondern in gewisser Weise um ihre Schöpfung. Eine der vielen Varietäten des Deutschen, Englischen, Französischen, Spanischen, Italienischen mußte ausgewählt, normiert und kodifiziert werden. In Italien, Vorreiter nicht nur in dieser Hinsicht, war es die *Academia della Crusca*, eine Gruppe von gelehrten Enthusiasten, die sich die Aufgabe, gute von schlechter Sprache zu scheiden, zu eigen machte und so das toskanische Italienisch in ihrem Wörterbuch von 1610 zum Standard erhob oder doch zu erheben suchte - nicht ohne erheblichen Widerstand vonseiten anderer Varietäten des Italienischen. In Frankreich war es der gnadenlose Richelieu, der zu diesem Zweck eine Akademie gründete und ihr die Aufgabe übertrug "(de) donner des règles certaines a notre langue" (Artikel XXIV ihrer Satzung vom 25. Februar 1635). Das Ergebnis, das die vierzig Unsterblichen in den folgenden Jahrzehnten zustandegebracht haben, war eher mager. Aber andere haben ihnen diese Aufgabe mit mehr Fleiß und Geschick abgenommen, und nur wenige Nationen haben die Idee der standardisierten Sprache für alle Angehörigen der Nation so rücksichtslos umgesetzt. In Deutschland wäre dies auch nur schwer möglich gewesen. Immerhin, 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 natürlich ein großes Wörterbuch der deutschen Sprache; daher die Anspielung auf die *Académie française*. Genutzt hat es wenig. Immerhin hat die Preussische Akademie der Wissenschaften, die Nachfolgeorganisation der Preussischen Societät, gut zweihundert Jahre später das von wagemutigen Verlegern und Gelehrten begonnene "Deutsche Wörterbuch" der Brüder Grimm unter ihre Fittiche genommen - ein Werk, das freilich ganz andere Ziele als die der Standardisierung verfolgte.

Das dritte wichtige Exempel, in dem - allerdings in einem anderen Sinne - die Standardisierung treibendes Moment hinter der Wörterbucharbeit ist, rührt aus der Beschäftigung mit bis dato schriftlosen Sprachen. Diese Beschäftigung war bis in die jüngste Zeit selten Selbstzweck, sondern sie sollte es möglich machen, die Bibel in die betreffende Sprache zu übersetzen. Dazu braucht man aber zuerst einmal ein Wörterbuch. Die so entstandenen, qualitativ sehr unterschiedlichen lexikographischen Beschreibungen - gleich ob zu Zwecken der Missionierung oder ob in politischem Auftrag beispielsweise von Kolonialverwaltungen - sind aber immer zugleich Kanonisierungen einer bestimmten Sprachform: in gewisser Weise schaffen sie erst eine Sprache, indem sie einen schriftsprachlichen Standard setzen und ihn damit de facto zur Norm erheben.

Das letzte und augenfälligste Beispiel einer Standardisierung über das Wörterbuch ist die Rechtschreibung. Die bis vor kurzem geltenden Regeln der deutschen Rechtschreibung stammen von Raumer und einigen anderen Germanisten aus der zweiten Hälfte des 19. Jahrhunderts. Man assoziiert sie aber stets mit dem Namen des Mannes, der aufgrund dieser Regeln eine Wortliste - eben ein Rechtschreibewörterbuch - zusammengestellt hat: Conrad Duden. Nach wie vor hat der "Duden" bei Gebildeten wie Nichtgebildeten den Status eines Heiligen Buches. Das meistverbreitete deutsche Buch ist ein Wörterbuch, ein Buch, in dem steht, wie es sein soll und nach dem man sich richten können möchte. Diesen normativen Wunsch nach Regelung sollte man nicht unterschätzen, wenn man sich daran macht, den Wortschatz einer Sprache zu beschreiben. Selbst den Intellektuellen scheint eine schier existentielle Verunsicherung zu befallen, wenn er nicht weiß, ob man *Schiffahrt* oder *Schifffahrt* schreibt.

2.2 Wißbegierde

Sehr spät erst ist der Gedanke aufgekommen, den Wortschatz einer lebenden Sprache einfach um seiner selbst willen

zu untersuchen und umfassend zu beschreiben. Zwar gibt es hier, wie immer, Vorläufer. Aber Bahn geschaffen hat sich diese Idee erst im Laufe des 19. Jahrhunderts, eng verbunden mit dem Aufkommen der historischen Sprachwissenschaft. [2] Den Anfang markiert das von Jacob und Wilhelm Grimm begonnene *Deutsche Wörterbuch*, dessen erste Lieferung 1852 erschienen ist und das danach noch 108 Jahre bis zu seinem Abschluß brauchte. Andere, wie das gewöhnlich OED genannte *New English Dictionary on Historical Principles* (1859 begonnen, 1928 abgeschlossen) oder das *Woordenboek der Nederlandsche Taal* (1852 von Matthias de Vries begonnen, 1998 von anderen abgeschlossen, Weltrekord an Bearbeitungszeit) schlossen sich an. Solche Wörterbücher sollen, wie schon zuvor entstandene Wörterbücher ausgestorbener Sprachen oder früherer Sprachstufen (etwa Raynouards *Lexique roman ou Dictionnaire de la langue des Troubadours* (1838 -1844) oder der schon erwähnte Graff (1834-42)), keinem der oben genannten praktischen Bedürfnisse dienen, noch dem der Standardisierung; es wollte einfach zusammentragen, was man über den deutschen, englischen, niederländischen, französischen Wortschatz in Vergangenheit und Gegenwart weiß. Wer darin liest oder etwas nachschlägt, tut dies, um seine Neugier oder seine Wißbegierde zu befriedigen. Daß das so dokumentierte Wissen auch praktischen Zwecken dienen kann, ist unbestritten; aber das war nicht der Grund, weshalb seine Schöpfer es in Angriff genommen haben.

Als Wissenschaftler ist man geneigt, solche wissenschaftliche Wörterbücher für das eigentliche Ziel der Lexikographie und für den Maßstab der Wörterbucharbeit zu halten. Sie sind aber die Ausnahme, ein schöner Luxus, den sich eine Gesellschaft leisten kann, die der Befriedigung des menschlichen Wissensdranges selbst einen hohen Wert beimißt. Sie sind aber auch möglicherweise indirekt von eminenter Bedeutung für die oben unter 2.1 und 2.2 genannten Gründe, sich mit dem Wortschatz einer Sprache zu befassen, weil sie dafür eine solide und verläßliche Grundlage schaffen.

Die drei hier genannten Motivationen der Wörterbucharbeit - praktische Bedürfnisse des Einzelnen, Standardisierung, reine Wißbegierde - schließen sich nicht aus, noch lassen sie sich scharf voneinander trennen. Sie überschneiden sich in vielfacher Weise, und das Ergebnis ist ein gewaltiges Spektrum von Wörterbüchern und Wörterbuchtypen, in die die Forschung vergebens Ordnung zu bringen versucht hat (vgl. die enzyklopädische Übersicht in Hausmann u.a. 1989 - 1991, 969 - 1573)[3]. In der allgemeinen Wahrnehmung assoziert man mit dem Begriff "Wörterbuch" sicher in erster Linie zweisprachige Wörterbücher und Rechtschreibwörterbücher; aber schon letztere verfolgen, wie der Rechtschreibeduden, neben der Information über die normgerechte Schreibweise noch andere Ziele, beispielsweise die Erklärung schwieriger Wörter.

Auch wenn die drei Motivationen einander nicht ausschließen, jedenfalls nicht im Prinzip, so stehen sie doch aus naheliegenden Gründen oft in einem gewissen praktischen Widerstreit, weil sich die unterschiedlichen Intentionen nur in Grenzen gemeinsam verwirklichen lassen. Es gibt kein Buch, das sich gleichzeitig als Reiseführer für japanische Touristen und als umfassende wissenschaftliche Dokumentation der Geschichte der deutschen Sprache eignet und darüber hinaus einen nennenswerten Beitrag zur Standardisierung leistet. Aber das schließt ja nicht aus, daß es andere Möglichkeiten gibt, diese auseinanderstrebenden Zwecke miteinander zu verbinden - vielleicht nicht vollständig, aber doch weithin. Ich glaube in der Tat, daß Digitale Lexikalische Systeme dies leisten können. Dazu ist es sinnvoll, sich zunächst einmal grundsätzlich zu vergegenwärtigen, was es denn überhaupt heißen soll, den Wortschatz einer Sprache zu beschreiben, was immer der Grund sein mag, weswegen man sich an diese Aufgabe macht. Was ist ein Wort, was sind die Eigenschaften von Wörtern, die man erfassen will, wie setzt sich der Wortschatz einer Sprache zusammen?

3. Wörter und Regeln

Zu den wenigen Dingen, über die sich die Sprachwissenschaftler einig sind, zählt die Annahme, daß jede Sprache ein *Lexikon* und eine *Grammatik* hat, d.h. eine Menge elementarer Ausdrücke, meist als "Wörter" bezeichnet, und eine Menge von Regeln, denen gemäß sich komplexere Ausdrücke aus einfacheren bilden lassen. Einige dieser Regeln bilden komplexe Wörter ("Morphologie"), andere aus mehreren Wörtern zusammengesetzte Ausdrücke ("Syntax"). Diese schon der Antike bekannten Unterscheidungen sind allerdings nicht immer klar. Zum einen ist der Begriff "Wort" nicht sehr gut definiert. Zum andern gibt es komplexe Ausdrücke, deren Bedeutung gut aus der Bedeutung ihrer Bestandteile ableitbar ist, während dies für andere nicht gilt. Erstere sind "kompositional", letztere

[2]Dies gilt übrigens auch für die großen Wörterbücher der klassischen Sprachen, insbesondere Griechisch und Latein. Vor Scheller (1783) und Passow (1831) waren sie - bei aller Vollständigkeit - doch vorrangig als Hilfsmittel zum Studium der Texte gedacht; freilich ist der Übergang vom "praktischen" zum "wissenschaftlichen", der Erforschung des Wortschatzes an sich dienenden Wörterbuch gleitend. Insbesondere kann ein für praktische Zwecke gedachtes Wörterbuch natürlich wissenschaftlichen Kriterien genügen.

[3]In den drei Bänden dieses monumentalen Werks, dem ich sehr viel verdanke, ist der Stand der Wörterbuchforschung bis zur Erscheinungszeit umfassend dokumentiert.

"lexikalisiert"; verwandte Bezeichnungen für diesen Gegensatz sind "produktiv" - "idiomatisch" und "freie - feste Verbindungen"; in jedem Fall ist die Unterscheidung ein Kontinuum. Die Lexikalisierung findet sich nur selten bei flektierten Wörtern, abgesehen vielleicht von Partizipialformen wie *entfernt*, *abgelegen* und ähnlichen. Überaus häufig ist sie hingegen bei zusammengesetzten Wörtern wie *Großmutter* oder *andrehen* und syntaktisch zusammengesetzten Ausdrücken wie *den Löffel weglegen* - wobei sehr viele dieser Ausdrücke sowohl eine kompositionale (oft "wörtlich" genannte) wie eine idiomatische Bedeutung haben. Gehören solche Lexikalisierungen zum Lexikon oder zur Grammatik einer Sprache? Darauf gibt es keine klare Antwort: ihrer Form nach sind sie komplex und regelbasiert, ihrer Bedeutung nach hingegen nicht, und die Übergänge sind fließend. Daher ist es sinnvoll, den Begriff "Lexikon" in einem etwas weiteren Sinn zu verstehen: es enthält alle elementaren Ausdrücke im engeren Sinne sowie jene Ausdrücke, die der Form, nicht aber der Bedeutung nach zusammengesetzt sind.

Das Lexikon einer Sprache existiert zunächst einmal in den Köpfen ihrer Sprecher, und in allen schriftlosen Sprachen existiert es nur dort; es ist in erster Linie ein "mentales Lexikon", oder, wenn man über Gruppen von Sprechern hinweggeht, eine Ansammlung von sich weithin überschneidenden, aber nicht gleichen mentalen Lexika. Ein Wörterbuch ist nur ein Versuch, bestimmte Aspekte solcher mentaler Lexika abzubilden. Wir wissen nicht genau, wie ein "mentales Lexikon" beschaffen ist (Dietrich 2002 gibt einen exzellenten Überblick über den Stand der Forschung). Einigkeit besteht aber darüber, daß es sich aus individuellen *lexikalischen Einheiten* zusammensetzt, zwischen denen bestimmte *lexikalische Relationen* bestehen. Der traditionelle Ausdrück *Wort* - statt lexikalische Einheit - ist zugleich zu weit und zu eng. Man würde den Ausdruck *schläfst* nicht gern als lexikalische Einheit ansehen, weil er eben nicht elementar, sondern regelbasiert ist; umgekehrt muß man zusammengesetzte Ausdrücke wie *Schuß vor den Bug* ("Warnung") oder *die Flinte ins Korn werfen* ("aufgeben") sinnvollerweise als lexikalische Einheiten betrachten, denn man kann sie bei der Beschreibung einer Sprache nicht in die Grammatik verweisen.

Man muß nun scharf unterscheiden zwischen einer lexikalischen Einheit und der Art und Weise, wie eine solche Einheit benannt wird. Wenn in einem Wörterbuch das Stichwort *Uhr* steht, gefolgt von allerlei Erklärungen, so ist diese Folge von Zeichen auf dem Papier natürlich nicht die lexikalische Einheit - es ist ein Name für diese Einheit, ein Name, der es dem Lexikographen möglich macht, sich auf die Einheit zu beziehen, und dem Nutzer, irgendwelche Informationen über die lexikalische Einheit nachzuschlagen.[4] Eine lexikalische Einheit selbst ist wesentlich abstrakter - es ist eine Verbindung von zumindest drei Bündeln von Eigenschaften:

1. Phonologische Eigenschaften: in diesem Fall ein glottaler Verschlußlaut (der im Deutschen nicht geschrieben wird, ein langer gerundeter vorderer Vokal, ein (in meinem Dialekt) uvularer Frikativ.
2. Semantische Eigenschaften: die lexikalische Einheit, hier ein einfaches Wort, wird dazu verwendet, um Objekte zu beschreiben, die dazu dienen, die Zeit zu messen.
3. Morphologische und syntaktische Eigenschaften; sie geben an, wie sich die betreffende lexikalische Einheit in komplexere Ausdrücke integrieren läßt: es gehört zur Flexionsklasse x, es verbindet sich mit dem Artikel *das*, usw. Falls die Einheit selbst zusammengesetzt ist, zählen dazu natürlich auch die Eigenschaften ihrer Zusammensetzung.

Diese drei Bündel von Eigenschaften sind normalerweise konstitutiv. Davon gibt es zwei Ausnahmen. Zum einen können - jedenfalls nach Annahme vieler Linguisten - sowohl die semantischen wie die phonologischen Eigenschaften fehlen, allerdings nicht gleichzeitig: es gibt phonologisch leere "Wörter", und es gibt semantisch leere "Wörter". Zum andern können an die Stelle der phonologischen auch andere "Gestalteigenschaften" treten, z.B. in Gebärdensprachen. Beides spielt für die folgende Diskussion keine besondere Rolle und wird daher nicht weiter berücksichtigt.

Auf der anderen Seite können sich mit den drei konstitutiven Eigenschaftenbündeln andere verbinden; die wichtigsten darunter sind graphematische - also jene Eigenschaften, die angeben, wie das "Wort" geschrieben wird, wenn es geschrieben wird: Uhr. Eine ganz andere Gruppe von Eigenschaften bezieht sich auf all das Sachwissen, das man mit der Kenntnis des "Wortes" verbindet, beispielsweise über die unterschiedlichsten Typen von Uhren, ihre Herstellung, ihre Geschichte usw. Im Prinzip ist dieses "enzyklopädische Wissen" etwas anderes als die Kenntnis der Wortbedeutung; was das Wort *Uhr* bedeutet, mag auch jemand wissen, der keine Ahnung von Geschichte oder Herstellung von Uhren hat. In der Praxis gestaltet sich diese Unterscheidung aber extrem schwierig, und dies ist folglich auch eines der größten Probleme für jeden, der die Bedeutung einer lexikalischen Einheit beschreiben will, sei diese nun ein einfaches Wort oder ihrerseits zusammengesetzt.

Schön wäre es nun, wenn es jeweils für ein Bündel phonologischer Eigenschaften ein festes Bündel morphosyntaktischer und semantischer gäbe. Das aber ist in keiner einzigen menschlichen Sprache auch nur annähernd der Fall. Der Selektionsdruck hat es so gefügt, daß es stets vielfältige Zuordnungen gibt. Ein und dieselbe Lautfolge (oder auch Graphemfolge) hat im Normalfall sehr unterschiedliche semantische und - dies vielleicht etwas seltener - unterschiedliche syntaktische Eigenschaften. Wer dies nicht glaubt, braucht nur einmal in einem gängigen

[4]Im Prinzip könnte man auch ganz andere Namen für lexikalische Einheiten wählen - man könnte sie z.B. durchnumerieren. Für Zwecke das Nachschlagens wäre das nicht sinnvoll, vielleicht aber für andere.

Wörterbuch nachzuschlagen, was alles an syntaktischen und semantischen Informationen unter dem Stichwort *legen* oder dem Stichwort *auf* steht. Diese "variable Zuordnung" ist nicht nur eines der auffällgsten Charakteristika menschlicher Sprachen, sondern vielleicht auch das größte Problem, wenn man versucht, ein Lexikon durch ein Wörterbuch oder in irgendeiner anderen Weise zu beschreiben; wir kommen darauf in Abschnitt 10 zurück.

Die Einheiten eines Lexikons stehen nicht isoliert da, sondern sie sind in vielfältiger Weise aufeinander bezogen; das Lexikon einer Sprache bildet immer ein komplexes Netzwerk von Einheiten, die in sich wiederum komplex sind. Die Relationen zwischen ihnen können an jeder der obengenannten Eigenschaften ansetzen. Die Einheiten können also phonologische Eigenschaften gemeinsam haben, sich beispielsweise reimen, die gleiche Silbenzahl oder das gleiche Akzentmuster haben. Sie können zum gleichen Flexionsparadigma zählen, zur gleichen Wortart oder die gleiche Rektion haben, d.h. bestimmte morphologische und syntaktische Eigenschaften gemeinsam haben. Der interessanteste Zusammenhang zwischen lexikalischen Einheiten gründet sich jedoch auf ihre semantischen Eigenschaften. Wenn man von "lexikalischen Relationen" spricht, so meint man im allgemeinen solche, die sich auf die Bedeutung beziehen: A kann annähernd gleichbedeutend mit B sein (Synonymie), beide können die entgegengesetzte Bedeutung haben (Antonymie), A kann ein Oberbegriff von B sein (Hyponymie), usw. All diese Relationen stecken voller Probleme, nicht zuletzt wegen der multiplen Zuordnung lexikalischer Eigenschaften (eine gute neuere Diskussion der wichtigsten semantischen Relationen findet sich in Murphy 2003).

Lexikalische Relationen im weiteren Sinne, also über die semantischen hinaus, sind für das mentale Lexikon sehr wesentlich; in zahlreichen psycholingustischen Experimenten ist nachgewiesen, daß die Aktivierung einer Einheit zugleich bestimmte andere Einheiten aktiviert ("priming"). Eine vollständige Beschreibung des Lexikons einer Sprache müßte dies im Grunde systematisch nachspielen. Das ist nie versucht worden. Es entspricht keinem praktischen Bedürfnis, es dient nicht der Standardisierung, und was die wissenschaftlichen Ambitionen angeht, so ist es zu schwierig. Dennoch sind solche Relationen in vielfältiger Weise in die lexikographische Beschreibung eingegangen. Das allertrivialste Beispiel ist die alphabetische Ordnung, die ja eine Relation zwischen bestimmten graphematischen Eigenschaften der lexikalischen Einheiten widerspiegelt. Sie findet sich allerdings nicht deshalb allenthalben, weil sie für das Lexikon selbst wichtig wäre, sondern weil man die Einheiten so besser finden kann. Sie ist eher eine Eigenschaft der Abbildung - des Wörterbuchs - als des abgebildeten Gegenstandes. Interessanter sind Reimwörterbücher, denen gemeinsame lautliche Eigenschaften zugrundeliegen, vor allem aber Synonymenwörterbücher, in denen versucht wird, den Wortschatz nach Bedeutungsverwandtschaft zu gruppieren, beispielsweise Roget fürs Englische oder Dornseiff fürs Deutsche. Alle Versuche in diese Richtung sind nur mäßig erfolgreich, auch wenn solche Wörterbücher durchaus einen gewissen praktischen Nutzen für all jene haben, die ihre Formulierung mal etwas variieren wollen. Das hat zwei Gründe. Der eine rührt aus dem Format: ein Wörterbuch und die von ihm erforderte zweidimensionale Repräsentation eignet sich nicht gut, um vielfältige semantische Relationen abzubilden. Der andere rührt aus der Natur dieser Relationen selbst: es ist sehr schwer, die Relationen zwischen auch nur einer kleinen Gruppe lexikalischer Einheiten anzugeben.

Dabei entspricht der Gedanke, die Bedeutung - und damit das, woran die meisten Wörterbuchbenutzer neben der Rechtschreibung interessiert sind - einer zentralen Idee der strukturellen Linguistik saussurescher Prägung: die Bedeutung eines Zeichens ist nicht durch das bestimmt, worauf es sich bezieht, sondern durch seine "oppositions" zur Bedeutung anderer Zeichen. Wenn das so ist, dann mußte sich auch die Bedeutung einer lexikalischen Einheit, also eines in semantischer Hinsicht elementaren Zeichens, vollständig über die Angabe der Relationen beschreiben lassen. Versuche in diese Richtung beschränken sich stets auf kleine, gut durchgegliederte thematische Ausschnitte eines Wortschatzes, beispielsweise die Bewegungsverben. Der erste wirklich substantielle Schritt darüber hinaus ist WordNet - ein von George Miller, Christiane Fellbaum und anderen entwickeltes Verfahren zur semantischen Analyse auf dem Computer (vgl. Fellbaum 1998); wir kommen noch darauf zurück. Die OULR-Darstellung des traditionellen Wörterbuchs eignet sich dafür sehr schlecht.

Im Voranstehenden habe ich ungefähr umrissen, was das Lexikon einer Sprache ist; es ist eine durchstrukturierte Menge von lexikalischen Einheiten, zwischen denen bestimmte Beziehungen bestehen. Wenn man ein solches Lexikon beschreiben will, sind folglich zwei Aufgaben zu lösen:

(a) Man muß die Eigenschaften der Einheiten angeben, die das Lexikon konstituieren, und
(b) Man muß angeben, was die Beziehung zwischen diesen Eigenschaften sind.

Gefiltert werden diese beiden Aufgaben zum einen durch den jeweiligen Zweck (etwa die in Abschnitt 2 genannten), zum anderen durch die Darstellungsmöglichkeiten, insbesondere eben durch das Format des Wörterbuchs. Bevor wir im nächsten Abschnitt auf die dadurch gegebenen Beschränkungen kommen, muß noch ein anderer Punkt angeschnitten werden: Was ist das Lexikon *einer Sprache*?

Gespeichert ist das Lexikon zunächst einmal in den Köpfen der Sprecher. Dort ist es nicht von Natur aus - es mag sein, daß bestimmte Komponenten sprachlichen Wissens angeboren sind, sicher aber nicht der Wortschatz -, sondern weil der betreffende Sprecher sich über Jahre hinweg ein gewisses lexikalisches Wissen angeeignet hat. Der Lexikograph ist aber im allgemeinen nicht daran interessiert, das lexikalische Wissen eines bestimmten Sprechers zu beschreiben, sondern den "Wortschatz einer Sprache". Nun muß man sich vor Augen führen, daß dies ein sehr eigentümliches Konstrukt ist. Es gibt nicht "den deutschen Wortschatz", weil es auch "die deutsche Sprache" nicht

gibt. Eine Sprache ist nichts, was gottgeschaffen über den Wolken thront und den Einzelnen unterschiedlich gut zugänglich wird. Was es gibt, ist das sprachliche Wissen und das entsprechende sprachliche Verhalten von Sprechern, die sprechen und hören, schreiben und lesen. Dieses Verhalten ist aber sehr uneinheitlichlich; es variiert nach Raum und Zeit, nach sozialer Schicht und kommunikativer Absicht, nach Redesituation und psychischem Zustand des Sprechers. Dies gilt für das Lexikon ebenso wie für die Grammatik. Es gibt viele Möglichkeiten, die damit gesetzten Spielräume der Variation einzuschränken, beispielsweise hinsichtlich der Zeit, hinsichtlich der schriftlichen Produktion bestimmter Autoren ("unsere besten Schriftsteller", "die gebildeten Schichten") und anderer Kriterien. Jeder Lexikograph definiert daher in gewisser Weise seinen "Varietätenraum" (Klein 1974) und beschreibt selektiv das lexikalische Wissen dieses Varietätenraums. Dieses Spektrum des Erfaßten wird zum einen von theoretischen Vorannahmen bestimmt (die deutsche Lernervarietät eines Gastarbeiters zählt nicht zum "Deutschen", der Dialekt von Mannheim hingegen schon), zum andern von praktischen Erwägungen, die vom Zweck der Beschreibung und von den Möglichkeiten des Formats abhängen: auch wenn man den Dialekt von Mannheim zum Deutschen rechnet, so wird man ihn doch nicht in den Ausspracheduden aufnehmen wollen.

Für eine lexikalische Beschreibung muß man daher nicht nur entscheiden, (a) welche Eigenschaften der lexikalischen Einheiten und (b) welche Relationen zwischen ihnen berücksichtigt werden, sondern (c) was denn der "Wortschatz" ist, den überhaupt man auszugsweise beschreiben will. Dies definiert die Aufgaben. Ihre Lösung hängt dann vom Zweck und vom Format ab.

4. Die Zwänge des Formats

Das traditionelle Format der Lexikographie ist das Wörterbuch. In welcher Weise beschränkt dieses Format die Lösung der drei genannten Aufgaben? Ich fange mit der dritten, der Entscheidung über den zu beschreibenden Wortschatz, an.

A. Auswahl der lexikalischen Einträge

Das bedeutendste deutsche Wörterbuch übehaupt, der "Grimm", enthält in seiner ersten Ausgabe etwa 300 000 Einträge, die als selbständige lexikalische Einheiten betrachtet werden können. Mehr findet sich in keinem deutschen Wörterbuch; aber es ist nur ein kleiner Ausschnitt des "deutschen Wortschatzes", so wie ihn die Brüder Grimm aufgefaßt haben. Die derzeit repräsentativste Zusammenstellung von Texten deutscher Sprache des letzten hundet Jahre, das Corpus des in Abschnitt 1 genannten *Digitalen Wörterbuchs der deutschen Sprache des 20. Jahrhunderts* (DWDS), enthält ungefähr 100 Millionen Wörter laufenden Textes von 1900 bis 1999 (siehe dazu im einzelnen Abschnitt 8 unten). Darunter finden sich etwa 8 Millionen *verschiedene Wörter*; nun sind dies zum Teil Flexionsvarianten und Eigennamen. Man wird aber selbst in diesem auf das 20. Jahrhundert begrenzten Corpus mit mehreren Millionen deutscher Wörter zu rechnen haben. Nun sind viele darunter kompositionale Zusammensetzungen wie *Bahnreise* oder *Parklücke,* die man vielleicht[5] nicht als lexikalische Einheiten rechnen will. Dies verringert die Zahl. Auf der anderen Seite sind aber alle syntaktisch komplexen, aber ihrer Bedeutung nach lexikalisierten oder halblexikalisierten Ausdrücke, wie etwa *zu Kreuze kriechen, zur Welt bringen*, *ins Zeug legen*, nicht berücksichtigt. Das erhöht die Zahl. 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 20. Jahrhundert veröffentlichten Texte vorkommen. Man braucht die Rechnung nicht fortzuführen. Kein Buch, keine Sammlung von Büchern ist in der Lage, diesen Wortschatz abzudecken. Dies sind die *Zwänge des Umfangs*.

Die zweite wichtige Einschränkung des Wörterbuchformats bei der Auswahl der lexikalischen Einheiten, die Berücksichtigung finden, ist die mangelnde Flexibilität. Scripta manent. Wenn ein Wörterbuch, oder auch nur ein Teil eines Wörterbuchs, erst einmal gedruckt ist, dann steht es zunächst einmal in dieser Form. Man kann nicht nach Belieben neue Einheiten hinzufügen, es sei denn - bei Mehrbändern - um den Preis der Uneinheitlichkeit, so wie sie für fast alle über viele Jahre laufende Wörterbuchprojekte charakteristisch ist. Dies sind die *Zwänge der mangelnden Erweiterbarkeit*.

B. Beschreibung der lexikalischen Eigenschaften eines Eintrags

---

[5]Ich sage "vielleicht", weil es beispielsweise für einen Lerner durchaus nicht durchsichtig zu sein braucht, wie sich die Bedeutung eines allem Anschein nach so durchsichtigen Wortes wie *Parklücke* aus den beiden Bestandteilen *Park-* und *Lücke* ergibt. Vielleicht denkt er, es ist eine Lücke im Park, oder eine Lücke in der dichten Bebauung, statt eine Lücke zum Parken.

Im Prinzip dieselben beiden Probleme ergeben sich, wenn es nicht um die Auswahl der lexikalischen Einträge, sondern um die Beschreibung ihrer Eigenschaften geht. Der Platz ist beschränkt, und was gedruckt ist, ist gedruckt. Ersterem versucht man durch Abkürzungen, Sonderzeichen, Querweise und andere Hilfsmittel abzuhelfen. Es ist erstaunlich, welche Erfindungsgabe hier die Lexikographie in ihrer langen Tradition an den Tag gelegt hat; man vergleiche hierzu beispielsweise Wiegand 1989, in der diese Möglichkeiten, wesentlicher Teil der "Mikrostrukur" eines Eintrags, mit großer Präzision und viel Liebe zum Detail beschrieben und auf den Begriff gebracht werden. Aber die Möglichkeiten sind begrenzt, und allzuviele Abkürzungen und dergleichen sind jedem Benutzer ein Greuel. Für letzteres, also die mangelnde Erweiterbarkeit, gibt es nur die Möglichkeit des Neudrucks oder des Nachtrags. Bei der Beschreibung der Eigenschaften gibt es darüber hinaus ein Gegenstück zum Zwang der mangelnden Erweiterbarkeit - nämlich die sehr beschränkten Möglichkeiten, falsche oder unzulängliche Informationen über diesen Eintrag zu ändern oder ganz herauszunehmen.

Es gibt aber noch ein drittes, viel grundsätzlicheres Problem aufgrund des Formats. Die einzelnen Eigenschaften - Phonologie, Morphosyntax, Semantik, eventuell alle weiteren Angaben - müssen schriftlich dargestellt werden; davon gibt es einige wenige Ausnahmen, etwa Bebilderungen; diese Möglichkeit ist aber gleichfalls aus Gründen des Formats sehr beschränkt nutzbar. Normalerweise jedoch benötigt man einen konventionellen, schriftlichen Code, im einfachsten Fall dieselbe Sprache, deren Lexikon das Wörterbuch selektiv beschreibt. Es gibt aber auch andere Codes. Für die phonologischen Eigenschaften muß man beispielsweise ein Schriftsystem ersinnen, von dem man hofft, daß es diese Eigenschaften akkurat wiedergibt, vor allem aber, daß der Benutzer ihn versteht und die rechte "Aussprache" daraus ableiten kann. Ich will dieses Problem - eigentlich eine ganze Klasse von Problemen, die sich für die einzelnen lexikalischen Eigenschaften ganz unterschiedlich gestalten - einmal als die *Zwänge des Schriftcodes* bezeichnen.

Wir haben diese Probleme hier nur für die drei - oder vier, wenn man die graphematischen mitrechnet - zentralen Eigenschaften einer lexikalischen Einheit erörtert. Sie werden erheblich komplexer, wenn man andere Angaben hinzufügen will, etwa über die Herkunft, die Entwicklung, den variierenden Gebrauch in verschiedenen Textsorten usw.

C. Beziehungen zwischen den Einträgen

Welche Einheiten immer man aufnimmt, wie immer man sie beschreibt - es sind auf jeden Fall viele, und damit kommt man zur dritten Aufgabe, nämlich den Zusammenhang zwischen ihnen darzustellen. Dazu gibt das Wörterbuch wenig her. Das Lexikon selbst, gleich in welcher Auswahl es beschrieben wird, ist stets eine vieldimensionale Struktur, definiert durch die Einheiten und die verschiedenen lexikalischen Relationen zwischen ihnen. Die platten Seiten eines Wörterbuchs mit ihrem Oben-Unten-Rechts-Links können diese Struktur sehr schlecht abbilden, ähnlich wie man eine Stadt schlecht durch die Beschreibung in einem Buch abbilden kann. Gemessen an einer Stadt ist das Lexikon einer Sprache, gleichwie die Auswahl daraus beschränkt ist, aber eine sehr komplexe, zudem abstrakte Struktur. 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 ist - gemessen an den Möglichkeiten der OULR-Darstellung - eine gute Suchstrategie. Sie stößt allerdings an gewisse Beschränkungen, wo man aus Gründen der Platzersparnis einzelne Einheiten zu den jedem Wörterbuchbenutzer vertrauten "Nestern" zusammenstellt (auch hierzu Wiegand 1989). Wesentlich gravierender ist schon, daß das Alphabetverfahren bei syntaktisch zusammengesetzten lexikalischen Einheiten in Schwierigkeiten gerät: wie ordnet man *den Teufel an die Wand malen* ein? Unter *Teufel*, unter *Wand*, unter *malen*, unter *an*? Vor allem aber faßt sie nicht zusammen, was 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 können, um nur eines der Probleme zu nennen, 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 *Zwänge der OULR-Darstellung* nur mit Instrumentarien, die mehrdimensionale Strukturen flexibel abzubilden erlauben.

Eine letzte Beschränkung des Wörterbuchformats, die ich hier kurz erörtern will, ist ganz anderer Natur. Umfassende Wörterbücher fügen den einzelnen Einträgen Beispiele für die Verwendung hinzu - entweder selbstgemachte oder Belege aus Texten irgendwelcher Autoren. Dies hat zunächst nicht unmittelbar etwas mit den drei oben genannten Aufgaben zu tun. Aber es macht es zum einen möglich, 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. Ich will diese Probleme des Wörterbuchformats einmal als *Zwänge der beschränkten Exemplifizierung* bezeichnen.

Eigentlich ist ein Wörterbuch ein sehr schlechtes Instrument, um das Lexikon einer Sprache in irgendeinem Ausschnitt darzustellen. Ich fasse die wesentlichen Beschränkungen noch einmal zusammen. Wir haben

- Zwänge des Umfangs
- Zwänge der mangelnden Erweiterbarkeit
- Zwänge des Schriftcodes
- Zwänge der OULR-Darstellung
- Zwänge der beschränkten Exemplifizierung.

Die Geschichte der Lexikographie ist zu einem großen Teil ein Kampf mit diesen Beschränkungen, die aus dem Format resultieren. Gibt es bessere Möglichkeiten? Nicht solange man sich auf das Format des gedruckten Buchs beschränkt.

5. Digitale Lexikalische Systeme

Die ersten Computerwörterbücher haben das tradierte Format nachgespielt: alphabetisch geordnete Listen von Wörtern mit ihrem Gegenstück in anderen Sprachen. Die digitalen Versionen klassischer Wörterbücher wie des OED, des Grimm, des Merriam-Webster bleiben im Prinzip gleichfalls in dieser Tradition; 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 all den im vorigen Abschnitt genannten Beschränkungen.

Digitale Lexikalische Systeme sind anders konzipiert. Wie ein Wörterbuch ist ein solches System eine selektive Abbildung des Lexikons einer Sprache. Anders ist es vor allem in dreierlei Hinsicht:

A. Es steht nicht auf dem Papier, sondern im Computer.
B. Datengrundlage - im wesentlichen ein Textcorpus mit geeigneter Verwaltung - und lexikalische Analyse sind konstant miteinander verknüpft.
C. Im übrigen läßt sich ein DLS durch vier gleichermaßen häßliche Schlagworte kennzeichnen: *Modularität, kumulative Entwicklung, inkrementelle Funktionalität, Methodenpluralismus.* Im Grunde sind es alles Abwandlungen ein und derselben Eigenschaft, nämlich einer hohen Flexibilität, was die Arbeit des Lexikographen und die Verwendung durch den Nutzer angeht.

Diese vier nicht schön benannten, aber überaus nützlichen Eigenschaften will ich nun im folgenden etwas näher erläutern.

5.1 Modularität

Das gesamte System besteht aus einer Reihe von Komponenten, die sich im Prinzip unabhängig voneinander bearbeiten und auch nutzen lassen. Dafür, wie man dies im einzelnen macht, gibt es 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 es also 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 Belegcorpus läßt dies relativ leicht zu.

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, insbesonder der "Standardvarietät" verknüpft sind.

Solche Module kann man sich in beliebiger Form vornehmen. Dabei muß man im Prinzip trennen zwischen

"Bearbeitungsmodulen" und "Nutzungsmodulen". Mit ersterem ist gemeint, daß ein Bearbeiter, oder eine Gruppe von Bearbeitern sich einen bestimmten abgrenzbaren Bereich vornimmt und bis zu einer gewissen Bearbeitungstiefe abschließt, beispielsweise die semantischen Eigenschaften der nicht zusammengesetzten Verben. Solche Einschränkungen sind aus praktischen Gründen sehr sinnvoll. Auf lange Sicht ("In the long run, we are dead", Lord Keynes) möchte man sicherlich nicht bloß eine solche Komponente haben, sondern ein Gesamtmodul "semantische Eigenschaften" (oder zumindest "semantische Eigenschaften" der Verben). Wesentlich ist jedoch, daß jedes Bearbeitungsmodul, sobald abgeschlossen, schon für sich sinnvoll nutzbar ist, auch wenn seine Funktionalität zunächst noch eingeschränkt ist.

5.2 Inkrementelle Funktionalität

Das klassische Wörterbuch ist eigentlich erst recht nutzbar, wenn es fertig ist. Das vor einem halben Jahrhundert begonnene Goethe-Wörterbuch, eine beeindruckende Leistung der deutschen Philologie, ist inzwischen beim Buchstaben G angelangt. Wer sich nicht dafür interessiert, was Goethe über *Gott*, sondern über *Welt* zu sagen hat, muß noch dreißig Jahre 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 durch den Olympier in 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 Corpus nicht auf einen Autor beschränkt ist. Man kann sich beispielsweise bei den syntaktischen Eigenschaften zunächst einmal mit den Ergebnissen eines "syntactic tagging" - d.h. einer vergleichsweise oberflächlichen Analyse nach Wortklassen oder auf eine Angabe syntaktischer Muster, in denen das Wort typischerweise vorkommt - begnügen. Auch das ist 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.

5.3 Kumulative Entwicklung

Damit ist das Gegenstück der inkrementellen Funktionalität auf Bearbeiterseite gemeint. Beim klassischen Wörterbuch erzwingt das Format eine Bearbeitung von A - 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.

5.4 Methodenpluralismus

Das klassische Wörterbuch muß für jeden Typ von Eigenschaften einheitlich vorgehen: man kann nicht für bestimmte Buchstaben mit einer grammatischen Analyse vom Typ A, sagen wir Head-driven Phrase Structure Grammar, und für andere mit traditioneller Schulgrammatik vorgehen. Ebenso kann man nicht die semantische Analyse für einen Teil über die vertrauten Paraphrasen der Bedeutung in normaler Prosa, für einen anderen hingegen über lexikalische Relationen wie Synonymie, Hyponymie usw. durchführen. In einem DLS kann man hingegen verschiedene Methoden miteinander kombinieren. So kann man beispielsweise für den gesamten abgebildeten Wortschatz mit den üblichen zu Listen geordneten Paraphasen bei der Semantik beginnen und dann bestimmte Bereiche mit anderen Methoden differenzieren. Manche lexikalische Felder, beispielsweise die Modalverben, die Farbadjektive oder die zeitlichen Konjunktionen, sind vergleichsweise gut durchstrukturiert, sodaß hier eine Analyse über Relationen oder, auch ein bekanntes Vorgehen, über semantische Merkmale sinnvoll und realistisch erscheint. Für andere Bereiche, insbesondere bei den Nomina, bietet sich ein solches Vorgehen nicht an. Es wäre ganz sinnlos, etwa die verschiedenen Teile eines Automotors über Synonymie- oder Antonymierelationen beschreiben zu wollen. Hier versagt auch die klassische Methode der Paraphrase 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.

Wie schon bemerkt, sind all diese Charakteristika letztlich nur verschiedene Ausprägungen der Flexibilität, die dadurch gewonnen wird, daß man sich von der zweidimensionalen Darstellung des Lexikon des klassischen Wörterbuchs lösen kann. Die Probleme eines DLS liegen nicht mehr in den Zwängen des Formats, sondern umgekehrt in der allzugroßen Freiheit. Es muß daher Sorge getragen werden, daß die Entwicklung und Ausarbeitung nicht ausufert. Dazu ist immer eine Koordinationsstelle nötig, die ein Grundkonzept mit bestimmten Freiheitsgraden entwickelt und seine Realisierung überwacht. In den folgenden Abschnitten will ich ein solches Konzept in seinen Grundzügen skizzieren. Es schließt an das schon erwähnte Vorhaben eines "Digitalen Wörterbuchs der Deutschen Sprache des 20. Jahrhunderts (DWDS)" an, das seit einigen Jahren an der Berlin-Brandenburgischen Akademie der Wissenschaften entwickelt wird (www.dwds.de und Klein 2004).[6] Ich bezeichne dieses Konzept im folgenden als "DLS der Akademie", kurz DLS-A.

6. Corpus und Module in einem DLS-A.

Du aber sitzt an deinem Fenster und erträumst sie dir.
Kafka

6.1 Die Grundlage

Wenn sich der Lexikograph an seine Arbeit macht, so muß er sie auf gewisse Fakten über das Lexikon, das er selektiv beschreiben will, stützen. Grundsätzlich steht das Lexikon einer Sprache im Kopf seiner Benutzer. Dort kann man nicht hineinsehen, und wenn, würde man nicht das Lexikon sehen, sondern bestenfalls die Zellen, in denen es gespeichert ist.[7] Man muß sich also anderer Methoden bedienen. Hier gibt es im wesentlichen drei Möglichkeiten, die einander nicht ausschließen:

A. Der Lexikograph stützt sich auf seine eigene Sprachbeherrschung, seine "Kompetenz".

B. Er stützt sich auf Beispiele für die Verwendung lexikalischer Einheiten, d.h. auf Belege, Belegsammlungen und Corpora.

C. Er stützt sich auf die Kompetenz anderer. Traditionell heißt dies, daß man nachschaut, was andere Lexikographen geschrieben haben. Man spricht hier oft von "sekundären Quellen" (Reichmann 1990, siehe auch Bergenholtz und Mugdan 1990). Es ist aber auch denkbar, daß man experimentelle Methoden anwendet, um zu testen, welche Eigenschaften die Sprecher einer Sprache mit bestimmten lexikalischen Einheiten verbinden. Dies geschieht selten, wenn man von der gelegentlichen Befragung von Experten absieht. Der Grund dafür liegt in dem großen Aufwand eines solchen Vorgehens; freilich beraubt man sich damit einer wichtigen Faktenquelle. Bei einem DLS ließen sich Informationen aus dieser Quelle wesentlich besser und flexibler nutzen, beispielsweise bei der Einschätzung des stilistischen Wertes oder der assoziativen Bedeutung, die eine bestimmte lexikalische Einheit hat.

All diese Quellen der Evidenz sind wichtig, alle haben gewisse Vorteile und Nachteile. Sich auf die eigene Kompetenz zu stützen, ist praktisch und einfacher, als sich auf die Analyse vieler Belege einzulassen; es ist aber auch unabdinglich, wenn man solche Belege untersuchen will. In der Praxis der Lexikographie arbeitet man zumeist mit einer Verbindung aller drei Methoden. Dabei kommt der Arbeit früherer Lexikographen oft ein sehr großes Gewicht zu: viele Wörterbücher, so wie sie auf den Markt kommen, sind Kompilationen aus anderen Wörterbüchern. Das ist nicht verwerflich, soweit es angegeben wird und ein gewisses Maß an eigener Leistung hinzukommt. Nur ein Narr kann das gesammelte Wissen der Altvorderen verschmähen und vollständig von vorn anfangen wollen. Aber dieses Wissen war begrenzt, oft sehr begrenzt: das ergibt sich aus den oben genannten Zwängen des Formats. Daher kann es nicht dabei bleiben.

---

[6] Entwickelt wurde das DWDS von einer kleinen Arbeitsgruppe, der seitens der BBAW Manfred Bierwisch, Wolfgang Klein (Leitung) und Hartmut Schmidt angehören. Eine Reihe von Experten, die bei wechselnden Gelegenheiten hinzugeladen wurden, hat wichtige Beiträge geliefert. Die eigentliche Arbeit wurde vor allem von Alexander Geyken (Leitung), Gerald Neumann und Ralf Wolz geleistet.

[7] Ich kann mir hier eine kleine Randbemerkung über die Relevanz der neuerdings so hochgehaltenen Hirnforschung für die Analyse der menschlichen Sprache nicht versagen. Eines ist das sprachliche Wissen, das in den Zellen des Gehirns gespeichert ist, ein anderes diese Zellen selbst. Der Sprachforscher ist an den Eigenschaften des sprachlichen Wissens interessiert. Dazu liefert die Analyse der Art und Weise seiner Speicherung wenig bei. Nirgendwo sieht man dies deutlicher als in der Lexikographie.

Eine lexikalische Beschreibung, die wissenschaftlichen Standards genügen soll, muß daher die Arbeit der Vorgänger heranziehen, die eigene Kompetenz ohnehin, sich aber im übrigen auf eine umfassende Untersuchung der Verwendung lexikalischer Einheiten stützen. Verwendung heißt dabei praktisch immer Verwendung in der aufgezeichneten Sprachproduktion: obwohl man ja Wörter nicht nur aktiv, sondern auch passiv gebraucht, gibt es zum Verstehen lexikalischer Einheiten nur wenige punktuelle Untersuchungen in der experimentellen Psycholinguistik; in der Praxis der Lexikographie wird die Sprachverstehensseite so gut wie nie berücksichtigt, sondern man nimmt stillschweigend an, daß Sprecherseite und Hörerseite des lexikalischen Wissens gleich, oder jedenfalls hinlänglich ähnlich sind.[8]

Das traditionelle Vorgehen ist die Exzerption. Aus einer mehr oder minder reichen Sammlung von Quellen - im Falle des "Grimm" rund 30 000 - werden von mehr oder minder kompetenten Lesern Belege mit mehr oder minder langem Kontext herausgeschrieben. Die Zettel werden alphabetisch zu Belegsammlungen geordnet, die dann die Grundlage der Bearbeitung bilden. Bei den größten bekannten Wörterbüchern liegt der Umfang eines solches Archivs bei bis zu fünf Millionen Belegen (etwa bei der zweiten Ausgabe des OED oder bei der Neubearbeitung des "Grimm" - dort allein für die Buchstaben von A - F). Dieses Vorgehen hat Vor- und Nachteile. Entscheidend ist zunächst einmal die Kompetenz der Exzerpierenden. Das OED beruht auf einer Initiative der *British Philological Society*, und so haben denn von 1859 an Hunderte von Mitgliedern dieser gelehrten Vereinigung, Lehrer und Liebhaber des Englischen, über mehr als zwei Jahrzehnte Belege aus der englischen Literatur herausgeschrieben, ohne daß ein einziger Artikel verfaßt worden wäre. Als die Philological Society im Jahre 1878 den schottischen Lehrer James Murray als Herausgeber anstellte, waren auf diese Weise rund zwei Millionen Belege zusammengekommen; aber der Bestand war sehr lückenhaft, ungleichmäßig und in vielen Fällen unbrauchbar, sodaß für einige Jahre über tausend neue "Leser" geworben wurden, die nun nach Murray gelegten Regeln eine weitere Exzerption vornahmen (Mugglestone 2000).

In gewisser Weise ist die Exzerption nie ein reines Datensammeln, sondern es ist bereits eine erste Bearbeitung, oft mit weitreichenden Folgen. Der Leser kann ja nicht jedes Wort herausschreiben, sondern er soll sich auf Verwendungen konzentrieren, die bisher nicht bekannt sind. In Murrays Worten:

> Make a quotation for *every* word that strikes you as rare, obsolete, old-fashioned, new, peculiar, or used in a peculiar way. (zit. nach Mugglestone 2000, 8)

Weder über die verschiedenen Leser hinweg noch für einen einzelnen Leser, da ja vielleicht über Jahre hinweg exzerptiert, garantiert dies eine hohe Konstanz. Die so getroffene Vorauswahl prägt die Analyse der Belege und damit das Wörterbuch in hohem Maße, auch wenn sie in gewissen Grenzen revidiert werden kann. Auf der anderen Seite ist so auch sichergestellt, daß der Lexikograph nicht im Material ertrinkt.

Seit rund drei Jahrzehnten treten zunehmend digitale Corpora an die Stelle des klassischen Belegarchivs, beispielsweise das *Collins COBUILD English Language Dictionary* (1987) oder dr siebzehnbändige umfangreiche *Trésor de la Langue Française* (1971ss)[9]. Solche Corpora erlauben eine unvergleichlich größere Breite der Abdeckung, aber auch eine gleichmäßigere, weniger von der Subjektivität der Exzerpierenden abhängige Aufarbeitung des Materials. Solche Corpora haben eigentlich nur ein Problem, und das ist der *embarras de richesse*. So beruht der Trésor auf einem Corpus von rund 90 Millionen laufender Wörter, allesamt Texte von 1789 bis etwa 1960; 70 Millionen davon entstammen literarischen Werken, die restlichen im wesentlichen der wissenschaftlichen Literatur. Aber das ist natürlich nur ein kleiner Ausschnitt der Texttypen, in denen sich der volle Wortschatz des Französischen niederschlägt. Ein großes Corpus ist nur sinnvoll, wenn es (a) gut ausgewählt ist, und, wichtiger noch, (b) durch gute Suchwerkzeuge erschließbar ist. Insbesondere müssen geeignete Filtermöglichkeiten vorgesehen sein, die es erlauben, schnell Wesentliches von Unwesentlichem zu trennen. In einem hundert Millionen umfassenden Corpus finden sich einige Millionen Belege des Wortes *und*. Es wäre natürlich, all diese Belege ansehen zu wollen. Freilich gibt es durchaus auch Zwecke, für die ein solch reiches Belegmaterial für ein einziges Wort sinnvoll ist, beispielsweise texttypologische und statistische Untersuchungen. Entscheidend für die Brauchbarkeit ist daher nicht der absolute Umfang, sondern die Zusammensetzung und die technische Aufarbeitung des Corpus.

In einem DLS ist die Datengrundlage, also im wesentlichen das Corpus, nicht festgeschrieben, sie läßt sich dynamisch weiterentwickeln. Man kann beispielsweise mit einem bestimmten Zeitraum beginnen und diesen dann systematisch ausweiten. Ebenso kann man mit bestimmte Texttypen beginnen, die man für besonders wichtig hält oder die auch besonders gut zugänglich sind, und sie schrittweise um andere ergänzen. Die klassische Lexikographie orientiert sich so gut wie immer an der geschriebenen Sprache. Dafür gibt es prinzipelle und praktische Gründe. Man

---

[8]Für den Sprachlerner trifft dies offenkundig nicht zu, ein Umstand, dessen man sich im Fremdspracherwerb sehr wohl bewußt ist. Aber dies hat meines Wissens nie nennenswerte Folgen für die konkrete lexikographische Analyse gezeitigt.

[9]Der Tresor liegt inzwischen auch in einer digitalen Version vor (frei zugänglich unter http://atilf.atilf.fr/tlf.htm), die in vielerlei Hinsicht bereits die Vorstellung eines Digitalen Lexikalischen Systems verwirklicht.

möchte beispielsweise die "geschriebene Sprache der besten Autoren" als Grundlage nehmen, und damit meint man im allgemeinen die gehobene Literatur, nicht die Art und Weise, wie der Pöbel in den Gassen redet. Aber selbst wenn man das will, ist es sehr schwierig, weil man dazu nur einen schlechten Zugang hat. Wir wissen nur sehr wenig darüber, wie die Mehrzahl der Leute in früheren Zeiten gesprochen hat, und auch für die Sprache der Gegenwart gibt es dafür nur ein sehr begrenztes Material. Aber natürlich ist die gesprochene Sprache in allen Kulturen primär: es ist die Sprache, die in einer Gesellschaft als erste entsteht - wenn es denn überhaupt eine geschriebene gibt -, die man als erstes lernt, und die man am meisten verwendet. Viele schreiben fast nie und lesen wenig. Die meisten Innovationen in einer Sprache gehen auf ihre gesprochene Form zurück; dies trifft nun zwar mehr auf Phonologie und Syntax zu - aber auch Veränderungen im Wortschatz sind nicht nur auf die geschriebene Sprache beschränkt. In einem umfassenden digitalen Corpus muß daher die gesprochene Sprache durch unterschiedliche Teilcorpora - die gesprochene Sprache ist ebensowenig einheitlich wie die geschriebene - repräsentiert sein.

6.2 Die Aufarbeitung

Ein Corpus, so gut es technisch erschließbar ist, ist noch keine lexikalische Analyse. Es müssen die in Abschnitt 2 genannten drei Aufgaben gelöst werden - nicht auf einen Schlag, sondern über einzelne Module. Mit welchen Modulen man beginnt, in welcher Analysetiefe man sie vorantreibt und welche man später wann hinzufügt, hängt vom Interesse der Beteiligten, vom Zweck der Arbeit und natürlich von den Arbeitsmöglichkeiten - und damit letztlich von der Finanzierung - ab. Für das DSL-A sind fürs erste fünf solcher Module geplant, von denen ich zwei in eigenen Abschnitten näher erläutern will. Die fünf geplanten sind:

A. Arbeitsmodul "Phonologische Eigenschaften"

Traditionell werden die phonologischen Eigenschaften - also kurz die "Aussprache" - in einer Lautschrift angegeben. Eine solche Lautschrift gibt aber selbst für jene, die sich damit auskennen, nur eine ungefähre Vorstellung der tatsächlichen Aussprache. Sehr viel sinnvoller wäre es, wenn man sich die Wörter tatsächlich anhören könnte. Mit den Möglichkeiten des Computers ist dies kein Problem. Im DLS-A sollen daher die lexikalischen Einträge per "sound-linking" mit dem Wortlaut verknüpft werden. Ich komme darauf in Abschnitt 8 ausführlich zurück. Der Nutzen dieses Vorgehens, das im übrigen eine Aufarbeitung der Lautseite nicht ausschließt, liegt auf der Hand, insbesondere wenn man an den Fremdsprachunterricht denkt.

B. Arbeitsmodul "Graphematische Eigenschaften"

Da die Schreibweise automatisch mit dem Text und damit mit dem Corpus gegeben ist, ist dieses Modul am unproblematischsten und bräuchte gar nicht als eigene Komponente erwähnt zu werden. Was aber so vergleichsweise einfach möglich wird, ist *eine vergleichende Analyse aller Schreibweisen* über das Jahrhundert hinweg, so daß (oder: sodaß ?) sich die Entwicklung der Orthographie von den frühesten Zeiten, die im Corpus durch Texte belegt sind, genau nachzeichnen läßt. Falls man das Corpus fortwährend für die Gegenwart fortführt, läßt sich auf diese Weise auch leicht ermitteln, wie sich beispielsweise die Rechtschreibreform von 1999 tatsächlich durchsetzt - bezogen immer auf die verfügbaren Texte. Eine solche Möglichkeit würde die wilde Spekulationen, die sich zu diesem Thema von den verschiedenen Parteiungen in die Welt gesetzt werden, durch solides Faktenwissen ersetzen - welche Konsequenzen immer man daraus ziehen mag (nicht als ob ich glaube, dass dieses Faktenwissen an den leidenschaftlich vertretenen Meinungen viel ändern würde).

C. Arbeitsmodul "Morphologie und syntaktische Eigenschaften"

Zu den morphosyntaktischen Eigenschaften, die bevorzugt in Wörterbüchern angegeben werden, zählen vor allem Flexionsparadigma, Wortklasse sowie einige Rektionseigenschaften. Zumindest erstere lassen sich weitgehend automatisch bestimmen; dies soll auch im DLS-A geschehen. Darüber gibt es eine Reihe weiterer Möglichkeiten, über die aber im Augenblick noch nichts feststeht.

D. Arbeitsmodul "semantische Eigenschaften"

Dies ist nach allgemeiner Auffassung der Kern der lexikographischen Analyse; es ist allerdings auch mit weitem Abstand der schwierigste Teil, zugleich jener, bei dem der Computer dem Forscher zwar die Arbeit erleichtern, aber kaum abnehmen kann: kein Computer kann sagen, welchen Bedeutungsbeitrag ein bestimmtes Wort zu der ganzen Konstruktion leistet, in der es vorkommt. Dazu muß man das Wort verstehen, und das kann einstweilen nur der

Menschmit hinlänglicher Brauchbarkeit[10].

Eine vollständige Aufarbeitung auch nur des derzeitigen Kerncorpus (siehe dazu unten Abschnitt 8) wäre utopisch. Im DLS-A sollen daher für die Anfangsphase zwei Beschränkungen gemacht werden, und zwar

(a) auf die "Grundwörter", d.h. jene, die nicht zusammengesetzt sind; dabei kann "zusammengesetzt" etwas unterschiedlich verstanden werden; dies sind jedoch Einzelfragen;

(b) auf Wörter, die besonders häufig vorkommen; hier lassen sich gewisse Zusatzkriterien anwenden (etwa: häufig in verschiedenen Texttypen oder nur in einem, usw.); auch dies sind Einzelfragen.

Auf diese Weise ließe sich eine Eingrenzung auf etwa 25 000 Wörter erreicht werden. Damit wäre immerhin der Kernbestand des deutschen Wortschatzes semantisch beschrieben. Eine künftige Fortführung könnte dann wiederum nach systematischen Kriterien diesen Bestand ausdehnen (beispielsweise alle Ableitungen von Verben oder bestimmte Komposita).

Damit ist nur etwas über die Eingrenzung gesagt, so wie sie derzeit im DLS-A geplant ist. Es ist klar, dass man sich auch ganz andere "Modularisierungen des semantischen Moduls" vorstellen kann, beispielsweise zunächst sämtliche Verben beschreiben, oder die Verwendungen innerhalb eines bestimmten Zeitraums - all dies entspricht dem Grundzug der kumulativen Entwicklung eines solchen Systems.

Zu welcher Eingrenzung man immer sich entscheiden mag - es ist damit noch nichts über die konkrete Methode der semantischen Analyse gesagt. Darauf wird in Abschnitt 10 näher eingegangen.

E. Arbeitsmodul "Kollokationen"

Zusammengesetzte lexikalische Einheiten, die mehr oder minder idiomatisiert sind - ich sage dafür im folgenden kurz zusammenfassend Kollokationen - wie *Aufmerksamkeit zollen, ins Abseits geraten, jemandem nicht grün sein*, *auf dem letzten Loch pfeifen* sind ein Stiefkind der traditionellen Lexikographie, allein schon deshalb, weil man sie in den Quellen schlecht finden und weil man sie, wie bereits in Abschnitt 3 gesagt, in den Artikeln eines gedruckten Wörterbuchs schlecht darstellen kann.[11] Ein DLS eröffnet in beiderlei Hinsicht ganz neue Perspektiven. Derzeit wird ein großer Teil der im derzeitigen DWDS-Corpus belegten Kollokationen in einem von Christane Fellbaum (Princeton) geleiteten Kooperationsprojekt aufgearbeitet und lexikalisch erschlossen (vgl. dazu den Beitrag von Christiane Fellbaum in diesem Heft).

7.3 Beides

Auch ein modulares Vorgehen verlangt seine Zeit. Zwar braucht man nicht mehr 106 Jahre zu warten, bis endlich auch die lexikalischen Einträge unter dem letzten Buchstaben bearbeitet und nachschlagbar sind, aber auch nur die 25 000 wichtigsten Wörter semantisch zu beschreiben, geht nicht von jetzt auf gleich. Anders als beim klassischen Wörterbuch ist es bei einem DLS möglich, sinnvolle Zwischenlösungen mit begrenzter Funktionalität einzuführen. Ein erster Schritt in diese Richtung sind intelligente Suchwerkzeuge für das Corpus, die es in gewissen Grenzen möglich machen, daß gleichsam ein jeder zum Lexikographen wird (sozusagen nach Josef Beuys' bekanntem Wort "Jeder Mensch ist ein Künstler"). Man denke etwa an einen Übersetzer, der ein bestimmtes, ungewöhnliches Wort - oder auch eine bestimme Kollokation - nicht kennt und für den es oft sehr viel hilfreicher ist, ein paar Vorkommen dieser lexikalischen Einheit im Corpus zu suchen, um es aus dem Kontext heraus zu verstehen, als die Bedeutung - zumal wenn bestimmte Konnotationen damit verbunden sind - in einem Wörterbuch nachzuschlagen. De facto spielen übrigens die Belegzitate in gedruckten Wörterbüchern oft genau diese Rolle: erst durch diese Verwendungen im Kontext erhält man ein Verständnis dafür, was eigentlich gemeint ist. Die Angaben in einem Wörterbuch sagen oft nicht so sehr, was die Bedeutung eines Wortes ist - sie sollten eher als eine Art Lernhilfe angesehen werden; wir kommen auf diesen Gedanken in Abschnitt 10 zurück.

In diesem Fall ist der Nutzer ohne weitere Hilfe auf seine eigene Interpretation angewiesen. Ein wesentlich darüber hinausgehender Schritt besteht darin, ein verfügbares Wörterbuch zu digitalisieren und mit dem Corpus zu verknüpfen. Dadurch ist es möglich, vom Wörterbuch ins Corpus und vom Corpus ins Wörterbuch zu springen. Anders gesagt, man kann sich zumindest die bisherige lexikographische Aufarbeitung zunutze machen, sie ergänzen und korrigieren, soweit das neue Belegmaterial des Corpus dies erlaubt.

Im DLS-A ist derzeit das Corpus mit dem "Wörterbuch der Deutschen Gegenwartssprache (WDG) " von

[10]Immerhin gibt es in der Computerlinguistik bereits eine ganze Reihe von Verfahren, Wortbedeutungen automatisch zu trennen (siehe den Überblick in Stevenson 2003); es ist aber fair zu sagen, daß sie nach wie vor in einem experimentellen Stadium sind.

[11]Eine gute Vorstellung von den Problemen vermitteln Cowie (1998) und Burger (2003).

Steinitz und Klappenbach (Berlin 1964 - 1977) verknüpft. Das WDG ist in mancher Beziehung überholt, aber es ist in der Klarheit seiner semantischen Analysen unübertroffen und damit eine der besten Grundlagen für eine lexikographische Aufbeitung der deutschen Gegenwartssprache. Vorgesehen ist die Verknüpfung mit dem "Etymologischen Wörterbuch der deutschen Sprache" (Pfeifer u.a. 1989). All dies illustriert aus der Warte des Nutzers das, was in Abschnitt 5.2 als "inkrementelle Funktionalität" bezeichnet wurde: man hat nicht alles, was man möchte, aber doch sehr viel, und für manche Zwecke genug.

Nach dieser allgemeinen Skizze gehe ich nun auf einige Komponenten des DLS-A etwas näher ein, im folgenden Abschnitt auf das Corpus des DWDS, das die Grundlage zum Corpus des DLS-A bilden soll, in den Abschnitten 9 und 10 auf Aussprache und Bedeutung.

8. Die DWDS-Corpusdatenbank

Ausgangspunkt der DWDS-Datenbank war der Plan, zwanzig Jahre nach Abschluß des WDG ein neues Wörterbuch der deutschen Sprache neuerer Zeit zu erarbeiten. Dabei war ursprünglich durchaus an ein konventionelles Wörterbuch gedacht. Es sollte die deutsche Lexik des ganzen zwanzigsten Jahrhunderts in möglichst repräsentativer Weise abdecken. Von Anfang an freilich war klar, daß die Datengrundlage kein Belegarchiv, wie noch beim WDG und allen anderen deutschen Wörterbüchern, sondern ein gut aufbereitetes digitales Corpus sein sollte. Nun kann es so etwas wie ein wirklich "repräsentatives" Corpus - und sei es auch noch so groß - im Grunde nicht geben. 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.

Die gesamte DWDS-Datenbank besteht aus zwei Corpora, einem KERNCORPUS und einem ERGÄNZUNGSCORPUS, die jeweils über geeignete Suchwerkzeuge erschließbar sind.[12] Letzteres umfaßt derzeit (Mitte 2004) etwa eine Milliarde Wörter laufenden Textes. Es ist dies ein opportunistisches Corpus, d.h. es sind Texte, die nicht nach wohlüberlegten Kriterien ausgesucht sind, 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 werden sie auch von verständnisvollen Verlegern umsonst zur Verfügung gestellt. Dies ist ja auch an verschiedenen Stellen bereits in der einen oder anderen Form geschehen. Der Aufschlußwert eines solchen Corpus für die lexikographische Analyse ist beschränkt, allein schon deshalb, weil solche Datenmassen vom Lexikographen nur schwer zu bearbeiten sind. Für viele statistische Zwecke und für selten belegte Wörter ist es aber durchaus von Nutzen.

Eigentliche Grundlage für die lexikographische Analyse ist jedoch das Kerncorpus, dessen Zusammensetzung systematisch geplant wurde. Das bezieht sich erstens auf den Zeitraum: es sollten in gleichmäßiger Verteilung Belege von 1900 bis 1999 erhoben werden. Eine solche Abgrenzung nach dem Kalender hat sicher etwas Willkürliches. Aber das gälte nicht minder für eine Abgrenzung nach "sachlichen", d.h. auf die lexikalischen Gegebenheiten zielenden Erwägungen, und so ist es vielleicht redlicher, sich direkt dazu zu bekennen. Wie immer - aufgrund der beliebigen Erweiterbarkeit eines digitalen Corpus spielt jede Abgrenzung dieser Art keine große Rolle: die Bestände lassen sich jederzeit in die Vergangenheit ausdehnen, und sie können jederzeit in die unmittelbare Gegenwart fortgeführt werden. Zweitens sollten auch unterschiedliche Textsorten in einigermaßen gleichmäßiger Verteilung berücksichtigt werden. Aufgenommen wurden Dokumente aus fünf Bereichen[13]:

A. Schöne Literatur; dazu zählt nicht nur die "gehobene" Literatur, die ja traditionell im Mittelpunkt der Wörterbucharbeit steht, sondern auch die sogenannte Trivialliteratur ebenso wie Kinderbücher; diese Texte machen 27%, also ein Viertel des Gesamtbestandes aus.

B. Journalistische Prosa, d.h. Zeitungen und Wochenschriften. Diese in sich wiederum recht heterogene Gruppe von Texten umfaßt mit 26% gleichfalls etwas mehr als ein Viertel des Bestandes.

[12]Die Texte in beiden Teilcorpora sind XML-codiert, linguistisch aufbereitet (d.h. lemmatisiert und mit Wortklassenangaben versehen) sowie durch verschiedene Metadaten beschrieben, sodaß eine sehr gute Recherche möglich ist. Verwaltet werden sie in einer *Oracle*-Datenbank.

[13]Rund 40 Millionen Wörter wurden als Bild- und Volltext eigendigitalisiert, und zwar größtenteils im sogenannten "Double-Keying"-Verfahren durch eine chinesische Firma nach in Berlin erstellten Arbeitsanweisungen; dabei werden die Texte unabhängig von zwei Schreibern abgetippt und anschließend verglichen; dies resultiert in einer Fehlerquote von unter fünf auf zehntausend Zeichen. Einige wenige Vorlagen konnten auch über optische Schriftkennung in Volltexte umgewandelt werden (für noch in Fraktur gedruckte Texte ist dies wegen der hohen Fehlerquote fast immer sinnlos). Alle weiteren Kerncorpustexte wurden eingeworben bzw. angekauft. Der Umfang der digitalisierten Texte entspricht rechnerisch ca. 1800 Monographien zu je 50.000 Textwörtern (dabei sind die Texte aus der gesprochenen Sprache abgerechnet).

C. Fachprosa (22 %), d.h. wissenschaftliche Texte, wobei ‚wissenschaftlich" nicht unbedingt im Sinne von wissenschaftlichen Originalveröffentlichen zu verstehen ist; so wurden beispielsweise sehr viele ins Gemeinverständliche tendierende Texte bedeutender Autoren wie Köhler, Einstein, Planck oder beispielsweise Friedrich Maurer für die Germanistik aufgenommen.

D. Gebrauchstexte (etwa 20 %); 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. Sie repräsentieren ein relativ breites Spektrum von lexikographisch verwertbarem Material.

E. (Transkribierte) Texte gesprochener Sprache (etwa 5%). Dies ist, wie schon bemerkt, für ein Corpus, das für lexikalische Zwecke gedacht wird, sehr ungewöhnlich - allein schon deshalb, weil es nur wenig Aufzeichnungen gesprochener Sprache aus der ersten Hälfte des 20. Jahrhunderts gibt. Das DWDS-Corpus enthält teils Parlamentsreden - zugegebenermaßen bereits eine etwas stilisierte Form gesprochener Sprache -, aber auch eine Reihe von Hörfunkreportagen aus dem Deutschen Rundfunkarchiv in Babelsberg aus den zwanziger und dreißiger Jahren.

Die Einteilung in fünf Texttypen sollte nicht überbewertet werden. Sie ist eher praktischer Natur. Da jedes Dokument durch bestimmte Angaben über Autor, Zeit, Art des Textes (z.B. Drama) gekennzeichnet ist, läßt sich der gesamte Bestand jederzeit für die spezifischen Zwecke "filtern"; so erhält man einen bestimmten Ausschnitt an Dokumenten, die man sich zu passenden Gruppen zusammenstellen und beispielsweise zeitlich ordnen kann.

Mit rund hundert Millionen Wörtern laufenden Textes ist das Kerncorpus nach wie vor beschränkt.[14] Immerhin, schon in seiner jetzigen Zusammensetzung reflektiert es die Lexik des Deutschen im 20. Jahrhundert in größerer Breite als jede andere Quellengrundlage. Dennoch repräsentiert es natürlich nach wie vor nur einen Ausschnitt des "gesamten deutschen Wortschatzes". Dies gilt zunächst einmal in historischer Hinsicht: lexikalische Einheiten, die nach Ausweis des Corpus im abgedeckten Zeitraum nicht mehr verwendet sind, werden nicht erfaßt. Dabei geht es oft weniger darum, daß bestimmte Wörter oder Wendungen *überhaupt* nicht mehr vorkommen, sondern, daß sie *in bestimmten Gebrauchsweisen* nicht mehr vorkommen. Zweitens werden nach wie vor bestimmte Arten von Texten nicht oder nur unzulänglich erfaßt, beispielsweise Werbetexte, die in lexikalischer Hinsicht oft besonders kreativ sind. Drittens werden bestimmte andere Dimensionen sprachlicher Variabilität, beispielsweise regionale Besonderheiten in der Verwendung nur unzulänglich abgedeckt.

In einem digitalen Corpus stellen all diese Beschränkungen der unvollständigen Abdeckung ein praktisches, aber kein prinzipielles Problem dar: ein solches Corpus ist dynamisch, es läßt sich jederzeit nach dem für ein DLS insgesamt geltenden Prinzip der kumulativen Erstellung erweitern. Im DLS-A ist hier vor allem eine Erweiterung auf Texte vor 1900 geplant und bereits in Arbeit.

Das hier beschriebene Corpus ist nicht einfach eine Zusammenstellung digitaler Texte - es ist durch Lemmatisierung und Wortklassenangaben relativ gut erschlossen. Außerdem ist es, wie schon erwähnt, mit dem "Wörterbuch der deutschen Gegenwartssprache" verknüpft. Dies reicht bereits für viele lexikalische Recherchen. Aber diese Verbindung ist natürlich erst die Keimzelle eines komplexen digitalen lexikalischen Systems, wie es oben skizziert wurde. Im folgenden gehe ich auf zwei der weiteren Arbeitsmodule etwas näher ein.

9. Die Aussprache

Am Anfang einer Sprache steht immer ihre gesprochene Form, und ein Linguist, wenn befragt, was denn ein sprachlicher Ausdruck ist, würde auf Anhieb zunächst immer sagen: eine Verbindung von Laut und Bedeutung. In Saussures berühmter Definition des sprachlichen Zeichens ist der *signifiant* selbstverständlich immer durch seine phonologischen Eigenschaften bestimmt. In der Lexikographie hat sich diese Denkweise nicht niedergeschlagen.[15]

---

[14]Übrigens ist es eine Illusion zu glauben, daß von einer bestimmten Textmenge an "ohnehin nichts mehr dazukommt". Wie Geyken (dieses Heft) aufgrund des Ergänzungscorpus festgestellt hat, steigt die Zahl der hinzukommenden Wörter bis zu einem Textumfang von einer Milliarde praktisch linear. Freilich handelt es sich bei den hinzukommenden Wörtern dann größtenteils um Komposita, an denen ja das deutsche sehr reich ist. Aber nicht alle dieser Komposita sind kompositional, d.h. nicht wenige davon wird man doch als lexikalisiert ansehen und demnach als lexikalische Einheiten behandeln müssen.

[15]Immerhin wird es im OED anerkannt; dort heißt es bereits im Vorwort zum ersten Band (S. XXXIV): "The pronounciation is the actual living form or forms of a word, that is, the word itself, of which the current spelling is only a symbolization" (zit. nach Ternes 41, 508). Natürlich ist auch die Aussprache nicht das Wort, sondern es ist nur ein Teil seiner Eigenschaften. Aber die Bedeutung dieser Eigenschaften wird zumindest gesehen, und

Angaben über die lautliche Seite eines Wortes (erst die einer zusammengesetzten lexikalischen Einheit) tauchen spät auf und sind zunächst sehr unsystematisch (siehe zur Geschichte und den grundsätzlichen Problemen Ternes 1989, 508 - 518). Anfangs beschränken sie sich auf Hinweise zur Betonung (so bereits beim Wörterbuch der *Academia della Crusca* von 1610). Das grundlegende Problem war es, eine geeignete Repräsentationsform, d.h. eine allgemein akzeptierte und verständliche Lautschrift zu finden. Solcher Lautschriften gibt es viele, die meisten davon handgestrickt. Erst mit der Gründung der *Association Phonétique Internationale* (API, auch IPA) im Jahre 1886 und der von ihr in den folgenden Jahren entwickelten Lautschrift (oft verkürzt auch API oder IPA genannt) wurde ein entscheidender Schritt getan. Heute stützen sich die meisten Wörterbücher auf dieses System, ganz gleich ob es sich um eigene Aussprachewörterbücher oder um zusätzliche Angaben, etwa in zweisprachigen Wörterbüchern, handelt. In den großen amerikanischen Wörterbüchern freilich hat man dazu bislang noch nicht durchringen können (vgl. die schöne Erörterung aus der Warte der lexikographischen Praxis bei Landau 2001, 118 - 127)

Welche Form im einzelnen nun gewählt wird - die Charakterisierung der lautlichen Eigenschaften einer lexikalischen Einheit durch eine Lautschrift hat immer mit drei großen Problemen zu kämpfen. Dies sind (a) die Variabilität und, damit einhergehend, die mangelnde Normierung der Aussprache, (b) die Fehleranfälligkeit von Lautschriften und (c) schließlich die Unzulänglichkeit der Repräsentation für den normalen Nutzer. Diese will ich nun kurz besprechen, und zwar jeweils bezogen auf die API-Lautschrift. Für andere Transkriptionssysteme gilt das Gesagte in verstärktem Maße.

A. Variabilität

Anders als in ihrer geschriebenen Form gibt es innerhalb einer Sprache gewöhnlich sehr viele sozial akzeptierte Aussprachevarianten. In Kiel spricht man anders als in Berlin, dort anders als in Köln oder Stuttgart, dort wiederum anders als in Wien, Zürich oder Bozen. Keine dieser Formen kann als *der* Standard gelten; dies gilt *mutatis mutandis* für alle Sprachen mit größerer Verbreitung. Welche Form also soll man in der Lautschrift wiedergeben? In der Regel beziehen sich die Lexikographen hier auf gewisse Konstrukte wie die vom Theodor Siebs konzipierte "Bühnenaussprache" (erstmals 1898), die aber in erster Linie für einen bestimmten Zweck, eben die verständliche Lautung auf der Bühne, gedacht war, die Idee der "received pronounciation" im englischen Englisch, oder auf die Aussprache, wie sie sich beispielsweise bei überrgionalen Rundfunk- und Fernsehprogrammen einpendelt. Wie immer man hier vorgeht - im Grunde sind alle Angaben dieser Art immer Vorschläge aus der Warte des Lexikographen, die allenfalls einen kleinen Teil der tatsächlich gesprochenen und unter den Betreffenden für gut befundenen Wirklichkeit darstellen. Niemand hat das Recht zu sagen, daß das in Hannover gesprochene Deutsch, das den Siebschen Kompromissen relativ nahekommt, gegenüber dem in Wien oder in Zürich gesprochenen Deutsch zu bevorzugen ist. Es gibt keine allgemein akzeptierte Norm, es wäre sogar schade, wenn es sie gäbe. Es gibt einfach viele Varianten, die alle ihr Recht haben.[16] Eine korrekte Darstellung müßte daher diese Varianten angeben, jeweils versehen mit Hinweisen, wann und wo sie realisiert werden. Diese Darstellung müßte sich, wie jede verläßliche lexikalische Analyse, auf empirische Untersuchungen stützen.[17]

B. Fehleranfälligkeit

Feler in der Orthographie fallen schnell auf, weil ein jeder die Orthographie einigermaßen beherrscht. Fehler in der Lautschrift werden normalerweise nur von ausgesprochenen Experten bemerkt. Auch braucht man von Anfang an solche Experten, um die Aussprache der einzelnen Wörter anzugeben, und ihrer gibt es nicht sehr viele. Deshalb sind

---

dementsprechend enthält das OED auch Ausspracheangaben, wenn auch in nicht sehr geglückter Form. Das steht in bemerkenswertem Gegensatz noch zum "Grimm". Allerdings ist im Deutschen die Aussprache auch wesentlich besser aus der geschriebenen Form zu erschließen als im Englischen.

[16]Es geht hier, wohlgemerkt, nicht um die "Dialekte", sondern um die ortsüblichen Realisierungen des Standards. Man kann sich als Beispiel vor Augen (bzw. vor Ohren) führen, wie es klingt, wenn ein Schwabe, ein Wiener, ein Züricher, ein Berliner, ein Aachener und ein Hamburger das *Lied von der Glocke* aufsagt. (Eigentlich möchte man sich allerdings gar nicht vorstellen, wie 'Brot und Wein' von Hölderlin gesprochen klingt).

[17]Was von der empirischen Grundlage der phonologischen Angaben selbst in ansonsten sehr guten Wörterbüchern zu halten ist, zeigt sich schön an einer der kuriosesten lexikographischen Fehlleistungen aller Zeiten - dem Eintrag "dord (dôrd), *n. Physics & Chem.* Density" in Webster's *Second International Dictionary* von 1934. Dord hatte versehentlich ein Bearbeiter einen Zettel mit dem Hinweis "D or d, Abkürzung für *density*" fehlgelesen. Das Bemerkenswerte daran ist nicht, es zu diesem erst fünf Jahre später entdeckten Fehler gekommen ist - es ist eher erstaunlich und bewundernswert, wie wenig derartige Fehler sich in solch umfangreichen Wörterbüchern finden - sondern daß der Bearbeiter in Klammern auch gleich die Aussprache dazugefügt hat. Wie sollte man es auch sonst aussprechen?

die Ausspracheangaben in vielen Wörterbüchern sehr fehlerbehaftet; nach den Untersuchungen von Max Mangold[18] nimmt die Zahl der Fehler unter dem allgemeinen Kostendruck, der mangelnden Verfügbarkeit von Experten und der Mühseligkeit der Aufgabe immer weiter zu. Dies ist natürlich kein grundsätzliches Problem, das mit dem Format des Wörterbuches zu tun hätte. Es ist aber ein Problem von großer praktischer Tragweite.

C. Unzulänglichkeit der Repräsentation

Eine Lautschrift wie die der API erlaubt im Prinzip eine sehr akkurate Wiedergabe lautlicher Eigenschaften. Aber dazu muß man sie sehr gut kennen, und nur die wenigsten Wörterbuchbenutzer haben nennenswerte Kenntnisse in Phonetik. Im allgemeinen lernt man im Fremdsprachenunterricht die wichtigsten Zeichen, dies zumeist am Beispiel einer bestimmten Sprache und daher mit entsprechender Einfärbung. Deshalb sind nur wenige Nutzer in der Lage, aus den lautschriftlichen Transkriptionen in einem Wörterbuch die tatsächliche Aussprache des betreffenden Worts richtig zu erschließen.

Zum andern stößt aber auch die beste Lautschrift an gewisse grundsätzliche Grenzen. Das Charakteristische an einer Aussprache schlägt sich oft in unterschiedlichen Graden in minimalen Diphtongierungen, in Graden der Stimmhaftigkeit bei stimmhaften Konsonanten, in leichten Unterschieden in der Aspiration nieder, nicht zuletzt auch in bestimmten suprasegmentalen Besonderheiten (manche Sprechweisen sind "singend"). Keine Lautschrift kann eine mehr als approximative Vorstellung davon vermitteln, wie man in Kiel oder in Berlin spricht, wenn man Hochdeutsch spricht. Dies ist nur möglich, wenn man es tatsächlich hört.

All diese Schwierigkeiten lassen sich zu einem großen Teil beheben, wenn man zur Darstellung der phonologischen Eigenschaften nicht eine Lautschrift verwendet, sondern die betreffenden lexikalischen Einheiten sprechen läßt und sie mit den graphematischen Eigenschaften verbindet.[19] In kleinerem Ausmaß ist dies schon bei einigen kommerziellen Wörterbüchern der Fall. Zu den lexikalischen Einheiten, deren Lauteigenschaften man so erfassen will, kann man hier auch zusammengesetzte Ausdrücke zählen. Jede Einheit läßt sich in der üblichen Weise suchen und dann anklicken. Falls erforderlich, kann man sie sie sich auch verlangsamt vorsprechen lassen; das ist für den Lerner eine wesentliche Hilfe. Damit ist das dritte Problem, das der unzulänglichen Representation, im wesentlichen gelöst. Die Fehleranfälligkeit ist geringer als bei einer Lautschrift, weil bei der Aufnahme jede einzelne Realisierung sofort abgehört und gegebenenfalls korrigiert wird. Das schließt Fehler nicht aus, macht sie aber weniger wahrscheinlich.

[18]Ich danke Max Mangold für ein höchst instruktives Gespräch in dieser Hinsicht.

[19]Dies schließt übrigens nicht aus, daß man in einem DLS zusätzlich eine lautschriftliche Transkription mitführt. Ein solches doppeltes Vorgehen ist keineswegs eine bloße Absicherung, so wie man den ersten Automobilen sicherheitshalber doch noch ein Pferd mitgegeben hat, sondern es gibt bestimmte Nutzungen, für die eine Lautschrift unabdinglich ist. So kann man auf dieser Basis sehr einfach ein Reimwörterbuch zusammenstellen; ebenso kann man Wörter nach der Aussprache suchen, wie das heute bereits in der digitalen Version des *Trésor de la Langue Française* (http://atilf.atilf.fr/tlf.htm) möglich ist. Ein dritter Nutzen lautschriftlicher Angaben liegt darin, daß sie eine exzellente Quelle für sprachstatistische Untersuchungen sind.

Es bleibt das Problem der Variabilität. Hier kann ein DLS wiederum die Möglichkeit des kumulativen Ausbaus und, damit verbunden, der inkrementellen Funktionalität ausspielen. Man kann sich, je nach Interesse und praktischem Zweck, für einige Varianten entscheiden und diese aufnehmen. Andere können nach Bedarf hinzugefügt werden. Für das DLS-A ist in der ersten Aufbaustufe vorgesehen, etwa 100 000 lexikalische Einheiten in norddeutscher, Wiener und Züricher Standardaussprache aufzunehmen, so wie sie etwa von Rundfunk- und Fernsehsprechern gesprochen wird. Weitere Varianten können nach Belieben hinzugefügt werden.[20]

Nicht gelöst ist das Problem der Normierung. Dies ist allerdings auch nicht die primäre Aufgabe eines wissenschaflichen Wörterbuchs. Auf der anderen Seite gilt allerdings, daß der normale Wörterbuchnutzer nicht einen großen Überblick über die verschiedenen Aussprachemöglichkeiten haben möchte, sondern an erster Stelle möchte er wissen, was "richtig" ist und wie er sprechen soll. Ein DLS erlaubt hier eine differenzierte Antwort: es kommt darauf an, wo und mit wem der Sprecher reden will, und je nachdem hat man seine Aussprache einzurichten. Es gibt keine Norm, es gibt viele Normen, und das ist schön so.

10. Die Bedeutung

10.1 Die fünf Plagen der Wortsemantik

Ein normaler Nutzer schlägt in einem Wörterbuch vor allem aus zwei Gründen nach: er möchte wissen, wie ein Wort geschrieben wird, oder er möchte wissen, was es bedeutet. Die rechte Schreibweise anzugeben, ist ein geringes Problem, jedenfalls unter dem Aspekt der Darstellung. Die Bedeutung anzugeben, ist eine fast unlösbare Aufgabe. Dafür gibt es prinzipielle und praktische Gründe

Was ist eigentlich eine Wortbedeutung? Darauf gibt es weder unter den Sprachwissenschaftlern noch unter den praktischen Lexikographen eine allgemein akzepierte Antwort. Sehr allgemein gesprochen sind es semantische Eigenschaften, die konventionell mit lautlichen und morphosyntaktischen Eigenschaften verbunden sind. Aber das gilt natürlich auch für zusammengesetzte Ausdrücke. Die Ausdrücke, um die es geht - eben die lexikalischen Einheiten - sollen in gewisser Weise elementar sein. Das kann sich nicht auf die Form beziehen, denn es gibt (vgl. Abschnitt 2) viele formal zusammengesetzte Einheiten - *sich in den Kopf setzen, zur Neige gehen, ins Gras beißen* -, die man gerne als semantisch elementar ansehen würde. Dies ist das erste große prinzipielle Problem der Wortsemantik, das der graduellen Lexikalisierung zusammengesetzter Ausdrücke. Welches Gewicht diesem *Problem der Idiomatizität* zukommt, kann man sofort deutlich machen, wenn man sich einmal anschaut, was ein gängiges zweisprachiges Wörterbuch unter dem Stichwort *legen* angibt; der bei weitem größte Teil eines Eintrags bezieht sich auf Zusammensetzungen wie *ins Zeug legen, in Ketten legen, Karten legen*, *Wert auf etwas legen.*

Das zweite liegt daran, daß die konventionelle Zuordnung nicht einheitlich ist. Das Wort *Verband* bedeutet etwas ganz anderes, je nachdem, ob man es im Vereinsrecht, in der Medizin, in der Mathematik oder bei der Flotte verwendet. Weniger offensichtlich, in Wirklichkeit aber massiver noch ist dieses Problem bei Wörtern wie beispielsweise *auf*. Ich gebe einfach ein paar Beispiele, die dies illustrieren:

(1) Die Tasse steht auf dem Tisch.
(2) Auf dem Regal standen etwa dreissig Bücher.
(3 Eva lebt auf dem Land.
(4) Horst arbeitet auf dem Rathaus.
(5) Auf der Decke sah man Reste des Frescos.
(6) Irgendwo auf dem Ball muß ein Preisschild kleben.
(7) Schneider hat eine Narbe auf der Fußsohle.
(8) Carlos lag auf den Knien/dem Bauch/dem Rücken.
(9) Auf dem Rothorn wehte ein eisiger Wind.

Allen unter (1) - (9) angegebenen Verwendungen ist gemeinsam, daß sie eine räumliche Konstellation zwischen zwei Entitäten beschreiben, zwischen einer Tasse und einem Tisch, einer Narbe und einer Fußsohle, dem Rothorn und dem Wind. Nur - diese räumliche Konstellation ist in jedem Fall offenbar eine andere. Gibt es überhaupt einen einheitlichen Bedeutungsbeitrag, den das Wort *auf* zur Gesamtbedeutung des Ausdrucks macht? Nach diesen wenigen

[20]Unter primär wissenschaftlichen Gesichtspunkten kann man hier nun noch einen Schritt weitergehen und genuine Dialekte, sagen wir die des Emmentals, aufnehmen. So läßt sich eine wesentliche Funktion von Dialektwörterbüchern in das System integrieren. Dies ist übrigens ein Gebiet, auf dem man die Unzulänglichkeiten der Repräsentation in einer gängigen Lautschrift besonders gut sehen kann. In meiner eigenen saarländischen Mundart ist die Auslautverhärtung stark eingeschränkt, aber nicht ganz verschwunden. Das Wort *dord*, wenn es es denn gäbe, würde eher /d□rd/ als /d□rt/ klingen, wie im Hochdeutschen - aber eben nicht ganz.

Beispielen möchte man es bezweifeln.[21] Dabei macht ein Blick in jedes etwas umfangreichere Wörterbuch sofort deutlich, daß dies nur eine kleine Auswahl ist, bei der zudem vier weitere wesentliche Verwendungen von "auf" bereits ausgeschlossen wurden:

- direktionale Verwendungen, wie *auf den Boden fallen*
- temporale Verwendungen, wie *auf der Betriebsfeier*
- zusammengesetzte Verben wie *(ein Buch, ein Ei, einen Ball) aufschlagen* usw.
- rein rektionsbedingte Verwendungen, wie *auf Hans warten, auf die Jungfrau vertrauen* , usw.

Traditionell wird das hier durch Beispiele angedeutete Problem unter Stichwörtern wie Mehrdeutigkeit, Polysemie und ähnlichen beschrieben: man hält die Form konstant und überlegt, welche semantischen Eigenschaften damit einhergehen können. Sinnvoller wäre es, allgemein von dem *Problem der variablen Zordnung* zu reden: zwischen verschiedenen sprachlichen Eigenschaftsbündeln kann es sehr verschiedene Zuordnungen geben; dies gilt nicht nur für Laut (oder Schrift) und Bedeutung, sondern auch für die verschiedenen syntaktischen Eigenschaften. Dies ist das zweite große Problem der Wortbedeutung.

Das dritte große Problem, bei dem man sich streiten kann, ob es eher ein prinzipielles oder ein praktisches ist, ist das Fehlen einer geeigneten Beschreibungssprache für semantische Eigenschaften. Jeder, der Deutsch kann, weiß, was das Wort *lachen* bedeutet; aber wie soll man es beschreiben? Ich gebe drei Beispiele, je eines für Deutsch, Englisch und Französisch:

lachen: durch eine Mimik, bei der der Mund in die Breite gezogen wird, die Zähne sichtbar werden u. um die Augen Fältchen entstehen, [zugleich durch eine Abfolge stoßweise hervorgebrachter, unartikulierter Laute] Freude, Erheiterung, Belustigung o.ä. erkennen lassen (nach Duden, Das große Wörterbuch der deutschen Sprache, 2. Ausgabe 1994)[22]

laugh: to give audible expression to an emotion (as mirth, joy, derision, embarassement, or fright) by the expulsion of air from the lungs resulting in sounds ranging from an explosive guffaw to a muffled titter and usu. accompagnied by movements of the mouth or facial muscles and a lighting up of the eyes (nach Webster's Third, Ausgabe 1981)

rire: éxprimer la gaieté par l'élargissement de l'ouverture de la bouche, accompagné d'expirations saccadées plus ou moins bruyantes (Petit Robert, Ausgabe 1972)

Man weiß nicht, ob man staunen oder das Wort anwenden soll. Die Präszision dieser Bedeutungsbeschreibungen ist bewundernswert. Aber die Sache würde natürlich sofort sehr viel klarer, wenn man einfach ein kleines Video zeigen könnte, in dem jemand lacht. In einem Wörterbuch ist dies freilich nicht möglich.

Das normale Instrument zur Bedeutungsbeschreibung ist eine natürliche Sprache - entweder jene, der die zu beschreibende lexikalische Einheit angehört, wie in den obigen Beispielen, oder eine andere (in welchem Falle sich die Beschreibung hier sehr einfach gestalten würde: **lachen** - to laugh, oder **lachen** - rire). Ob lang oder kurz, eine Bedeutungsbeschreibung ist traditionell eine *Paraphrase.* Und da eine lexikalische Einheit aufgrund der variablen Zuordnung in aller Regel mehr als eine Bedeutung hat, hat die normale Beschreibung der semantischen Eigenschaften in einem Wörterbuch gewöhnlich die Form einer bis zu einem gewissen Grade strukturierten Liste; man reiht die verschiedenen Bedeutungen nicht nur auf, sondern versucht sie nach Ähnlichkeit, nach ihrer historischen Entwicklung oder sonstigen Kriterien ein wenig zu ordnen. Dieses Vorgehen ist so vertraut, daß man es sich eigentlich gar nicht anders vorstellen kann. Es hat aber mehr als eine Crux.

Die wichtigste davon ist, daß die Bedeutungsbeschreibung oft weniger gut zu verstehen ist als das zu beschreibende Wort - jedenfalls in einsprachigen Wörterbüchern. Samuel Johnson hat sich in seiner berühmten Definition des Begriffes Netzwerk darüber lustig gemacht:

Network: Any thing reticulated or decussated, at equal distances, with interstices between the intersections.

Wie die obigen Beispiele für *lachen*, insbesondere das aus Webster's, zeigen, ist dies der Realität durchaus nicht fern: es ist eine *explanatio obscuri per obscurius*, die vor allem dann kurios wird, wenn das Explanandum nicht sehr obskur ist. Dieses *Problem der Beschreibungssprache* ist den Lexikographen natürlich mehr als vertraut (vgl. dazu die ganz aus der Praxis kommenden nüchternen Überlegungen von Landau 2001, 153 - 216). Eine denkbare Lösung besteht darin, zur

[21]In der Neubearbeitung des Grimmschen Wörterbuchs beruht der Artikel *auf* auf insgesamt 17 000 Belegen, die bis ins Jahr 1500 zurückreichen und ein noch viel weiteres Spektrum an Bedeutungen belegen.

[22]Hier zum Vergleich, was Adelung (1777ss) schreibt: *eine angenehme und durch Lust erregte Erschütterung der Nerven durch Verlängerung und Öffnung des Mundes, und zuweilen auch durch einen damit verbundenen inartikulierten Schall an den Tag legen.*

Definition selbst nur "einfache Ausdrücke" zu verwenden, also nicht von *mirth*, *guffaw* oder *titter* zu reden, wenn man das Wort *laugh* erklärt. Dies ist ein schöner Gedanke, er läßt sich nur nicht durchführen, wenn man über eine erste, oberflächliche Erklärung, wie sie vielleicht für Lernerwörterbücher angebracht ist, hinauskommen will.

Das vierte schwierige Problem läßt sich durch die Frage umschreiben: Wie weiß der Lexikograph eigentlich, welche semantischen Eigenschaften mit einer bestimmten Lautfolge (oder Buchstabenfolge) verbunden sind? Die Kenntnis der Bedeutung eines Wortes ist nicht angeboren; sie steht als Ergebnis eines komplexen Lernprozesses im Kopf. Dazu gibt es keinen direkten Zugang - man kann die semantischen Eigenschaften eines Wortes nur über seinen Gebrauch in der Kommunikation ermitteln.[23] Wie aber geht dies? Normalerweise so, daß man eine Lautfolge hört (bzw. eine Buchstabenfolge liest) und aus den Informationen, die aus dem Kontext zugänglich sind, Rückschlüsse darauf zieht, welchen Beitrag diese Lautfolge zur Gesamtbedeutung der Äußerung im Kontext macht. Die Kontextinformationen können drei Quellen entstammen: dem Weltwissen, der Redesituation und dem unmittelbaren sprachlichen Kontext, also dem, was vor oder nach der zu analysierenden Wortfolge kommt. Das ist schwierig, und bis heute hat niemand recht verstanden, wie es Kindern im Vorschulalter gelingt, dies mit einer Geschwindigkeit von durchschnittlich 5 - 8 Wörtern pro Tag zu schaffen (vgl. Clark 1993). Die Hauptschwierigkeit liegt darin, daß die kontextuellen Informationen, so reich sie auch sein mögen, die relevanten semantischen Eigenschaften oft hinlänglich einschränken; sie können es auch gar nicht, denn sonst wäre der Beitrag des zu analysierenden Wortes ja überflüssig. Man kann sich dies gut an der obigen Definition von *laugh* aus Webster's Third vor Augen führen: nur die wenigsten Leser werden wissen, was *guffaw* bedeutet. Man kann sich aufgrund des Kontextes eine gewisse Vorstellung machen - es ist wohl eine Art Geräusch, aber was genau, läßt sich aus diesem Kontext allein nur sehr unzulänglich erschließen.[24] Also braucht man weitere Kontexte; aber hier rennt man sofort in das weiter oben genannte Problem der variablen Zuordnung: eine Lautfolge kann sich mit sehr unterschiedlichen semantischen Eigenschaften verbinden. Was der Lexikograph darstellt, ist daher immer das Ergebnis eines Lernprozesses, der über viele Kontexte generalisiert. Insoweit er die Sprache gut beherrscht, ist dieser Lernprozeß sehr fortgeschritten. Schwierig wird es vor allem dort, wo es um geringfügige Abweichungen geht, beispielsweise in der historischen Entwicklung. Im Germanistikstudium lernt man, daß das Wort *zwivel* zu Beginn des "Parzifal" eben nicht "Zweifel" bedeutet: *Ist zwivel herzen nachgebur, das muoz der sele werden sur*. Irgendwie ist es mit Zweifel verwandt, aber was genau nun gemeint ist - da scheiden sich die Geister der Kundigen. Nun ist dies vielleicht ein extremer Fall, aber im Prinzip hat man bei jeder lexikalischen Einheit das *Problem des indirekten Zugangs* aufgrund von variablen und unterdeterminierenden Kontextinformationen.

Die letzte Schwierigkeit, gleichfalls ein grundsätzliches wie praktisches Problem der Bedeutungsanalyse, ist die Scheidung zwischen den semantischen Eigenschaften der lexikalischen Einheit und dem ganzen Sachwissen, die sich, von Sprecher zu Sprecher variierend, mit dieser Einheit verbindet. Wir haben dies bereits kurz in Abschnitt 2 am Beispiel der Einheit *Uhr* erwähnt. Man würde nicht annehmen, daß sich der Beitrag, den *Uhr* zur Bedeutung einer Äußerung macht, ändert, wenn man lernt, daß es auch Quarzuhren gibt. Sonst müßte man die Vorstellung eines eine größere Gruppe von Sprechern übergreifenden Lexikons ganz aufgeben, denn das Sachwissen der Einzelnen schwankt natürlich. Im konkreten Fall ist die Trennung aber sehr schwierig. Im WDG heißt es unter **backen**: *eine unfertige Speise der Ofenhitze aussetzen, so daß sich eine Kruste bildet und sie eßbar wird.* Dies ist eine schöne, klare und gut nutzbare Bedeutungsbeschreibung. Aber natürlich kann man auch im Mikrowellenherd einen Kuchen backen; da bildet sich aber keine Kruste (in der Frage der Eßbarkeit der so bereiteten Speise kann man geteilter Meinung sein). Hat sich durch die Erfindung des Mikrowellenherds die Bedeutung des Wortes *backen* geändert? Es ist fast unmöglich, dieses Problem ohne eine gewisse Willkür zu lösen.

Wir haben damit fünf fundamentale Probleme der semantischen Analysie lexikalischer Einheiten genannt:

- dass Problem der Idiomatizität
- das Problem der variablen Zuordnung
- das Problem der Beschreibungssprache
- das Problem des indirekten Zugangs
- das Problem der Abtrennung des Sachwissens.

Sie sind die fünf biblischen Plagen der Lexikographie und auch, soweit sie sich denn überhaupt ernsthaft mit Fragen der Wortsemantik beschäftigt, der theoretischen Linguistik. Ihre Bewältigung wird durch die in Abschnitt 3 diskutierten Zwänge des Wörterbuchformats sehr erschwert. Lassen sie sich durch ein Digitales Lexikalisches System aus der Welt schaffen?

Nein, denn die Probleme sind grundsätzlicher Natur. Aber man kann aufgrund der hohen Flexibilität eines solchen Systems einer Lösung näherkommen. Vor allem aber kann man sich etwas zunutze machen, das auch sonst die Grundlage aller lexikalischen Kenntnisse ist - nämlich das Lernen aus dem Kontext. Das will ich an fünf Punkten

---

[23]Dies besagt übrigens nicht, daß die Bedeutung der Gebrauch *ist*; selbst Wittgenstein, dem diese Auffassung oft zugeschrieben wird, drückt sich sehr vorsichtig aus (vgl. *Philosophische Untersuchungen* §43).

[24]Es heißt "wieherndes Gelächter", ich habe es nachgeschlagen.

erläutern.

10.2 Visuelle und auditive Darstellungsformen

Das klassische Wörterbuch arbeitet mit Paraphrasen, die lang oder kurz sein können und derselben oder einer anderen Sprache entstammen. Es gibt aber durchaus zumindest für manche Bereiche des Lexikons andere Möglichkeiten der Darstellung, die neben das klassische Vorgehen treten und es in manchen Fällen auch vollständig ersetzen können. Die wichtigsten dieser Möglichkeiten sind visuelle und auditive Darstellungen, von denen erstere bereits in "Bildwörterbüchern" genutzt werden. Dort gibt es allerdings erhebliche Einschränkungen aufgrund des Formats; sie bilden daher, außer im Bereich des technischen Übersetzens, die Ausnahme. In einem DLS kann man sie systematisch ausbauen, zum einen, indem man einfach mehr und differenziertere Abbildungen hinzunimmt. Dies gilt nicht nur für den technischen Bereich, sondern für viele Bedeutungen, die sich auf physische Objekte beziehen, sagen wir *Hand, Daumen, Apfel, Katze*. Man kann aber auch bewegte Bilder nutzen. Wir haben oben schon erwähnt, daß sich die Bedeutung von *lachen* statt durch hochkomplexe und schwer verständliche Paraphrasen in den verschiedenen Sprachen sehr viel einfacher durch einen kleinen Film erläutern ließe, in dem jemand lacht: ein Film sagt mehr als tausend Worte. Solche Filme lassen sich mit anderen aus demselben Bedeutungsfeld (*lächeln*[25]*, weinen, schreien, wimmern*) verbinden. Ein anderes Beispiel sind Bewegungsverben: es ist schier unmöglich, die semantischen Eigenschaften von *humpeln* und *hinken* so durch Paraphrasen zu beschreiben, daß der Unterschied zwischen ihnen deutlich wird. Aber zwei kleine Filme machen es deutlich. Etwas eingeschränkter, aber in manchen Fällen doch sehr sinnvoll ist die Darstellung der Bedeutung über auditive Informationen, beispielsweise bei Verben wie *bellen, zwitschern, keuchen, seufzen*.

Was man sich in all diesen Fällen zunutze macht, ist die Idee des *lexikalischen Lernens*, auf dem letztlich jedes Wissen über die semantischen Eigenschaften von Wörtern beruht. Dabei ist man natürlich auch mit den oben genannten Problemen dieses Lernens konfrontiert: der Kontext ist nicht immer restriktiv genug, und die Bedeutung schwankt in verschiedenen Kontexten. Auch hier kann man in flexibler Weise den Prozeß des lexikalischen Lernens nachspielen, indem man - gut ausgewählt - verschiedene Repräsentationen aufnimmt, also verschiedene Arten des Lachens oder verschiedene Äpfel. Noch einen Schritt weitergetrieben, läßt sich dies zu einem "lexikalischen Lernsystem" ausbauen, das man beispielsweise für den Zweitspracherwerb nutzen kann.

10.3 Darstellung durch die Beziehung zu andern Einheiten

Die Einheiten eines Lexikons stehen in bestimmten semantischen Relationen zueinander. Nicht jeder teilt die oben erwähnte Auffassung des klassischen Strukturalismus, nach der sich die Bedeutung eines Wortes allein aus diesen Relationen ergibt. Aber zumindest wesentliche Teilaspekte lassen sich über Bedeutungsrelationen charakterisieren - ganz abgesehen davon, daß dieses Netz ein wesentlicher Zug eines jeden Lexikons ist und daher einen wesentlichen Bestandteil seiner Beschreibung ausmachen sollte. Im Format des Wörterbuchs geht dies sehr schlecht, und es wäre vielleicht auch für den normalen Benutzer von begrenztem Wert. In einem DLS ist es aber möglich. Es gibt zumindest zwei größere lexikographische Vorhaben, in denen dies nach wie vor ausschnittweise, aber doch für erhebliche Teilbereiche des Lexikons durchgeführt wurde. Dies sind das von George Miller, Christiane Fellbaum und anderen seit Mitte der Achzigerjahre in Princeton entwickelte System WORDNET, das auf Relationen aufbaut, und das von Charles Fillmore und anderen Ende der Neunzigerjahre in Berkeley System FRAMENET, in dem die Bedeutung von Wörtern durch ihre Stellung in sogenannten "frames" - das ist das gebündelte Wissen zu einem bestimmten inhaltlichen Bereich - erfaßt wird (siehe hierzu Fellbaum 1998 und Fillmore u.a. 2003). Diese Systeme sind hochentwickelt; beide wurden zunächst am Beispiel des Englischen entwickelt, inzwischen auch auch auf eine Reihe anderer Sprachen angewandt. Sie eröffnen ganz neuartige Einsichten in die Struktur eines Lexikons. Allerdings sind schlecht geeignet, wenn man nur einmal schnell nachschlagen will, was die Bedeutung eines Wortes wie *Verband* ist und wie dieses Wort in verschiedenen Kontexten verwendet wird. Es ist aber nicht so, als wäre die Hinzunahme von semantischen Relationen in die Beschreibung nur aus wissenschaftlichen Gründen interessant. Sie ist, um nur zwei praktische Nutzungen zu nennen, für die Computerlinguistik hilfreich, und sie bildet die Basis von "Synonymenwörterbüchern" - d.h. man kann beispielsweise aus Gründen der stilistischen Variation oder auch der Präzision des Ausdrucks nach "bedeutungsverwandten" Wörtern suchen. Der Vorzug eines DLS besteht eben darin, daß verschiedene Repräsentationsweisen zu ganz verschiedenen Zwecken nebeneinander bestehen können.

10.4 Rasche Korrektur und Ergänzung

In einem klassischen Wörterbuch ist die Bedeutungsbeschreibung festgelegt, bis die nächste Ausgabe erscheint; in

[25] "*durch eine dem Lachen ähnliche Mimik Freude, Freundlichkeit o.a. erkennen lassen*", Duden, *Das große Wörterbuch der deutschen Sprache*, Ausgabe 1994, zu *lächeln*.

einem DLS kann sie jederzeit geändert und ergänzt werden. Diese Flexibilität erlaubt aber nicht nur punktuelle Korrekturen, sondern man kann sie nutzen, um aus einem bestehenden, noch nach Umfang, Abdeckung und Analysetiefe sehr beschränkten Wörterbuch eine sehr viel anspruchsvollere Beschreibung aufzubauen. Weiter oben in Abschnitt 5.3 wurde erwähnt, daß die Corpusdatenbank des DWDS bereits jetzt mit der digitalen Variante des "Wörterbuchs der deutschen Gegenwartssprache" vernetzt ist: man kann jederzeit von einem Eintrag im Wörterbuch ins Corpus springen und nach einschlägigen Belegen suchen. Nicht alle Einträge sind dort belegt; dies ist eine Frage der Corpuszusammensetzung. Viel wichtiger ist aber, daß es viele lexikalische Einheiten im Corpus gibt, die entweder gar nicht oder - mutmaßlich der wichtigere Fall - nicht in einer bestimmten Verwendung im WDG vorkommen, und dies ist eine Unzulänglichkeit des WDG. ImWDG finden sich rund 100 000 lexikalische Einheiten, aber allein im Kerncorpus des DWDS gibt es mehrere Millionen lexikalischer Einheiten. Daß dies keine Besonderheit des WDG ist, zeigt der Beitrag Geyken in diesem Heft.

Man kann nun das WDG als Ausgangspunkt für eine wesentlich umfassendere Beschreibung nehmen. Dazu werden als erster Schritt unter jeden Eintrag des WDG die entsprechenden Belege in passend geordneter Form einsortiert (etwa nach Texttypen). Für nicht vorhandene Einträge wird ein neues Stichwort angelegt. Diese Neuzugänge kann man alphabetisch oder nach einem sonst für die Bearbeitung zweckmäßigen Kriterium ordnen (z.B. nach absoluter Häufigkeit des Vorkommens, nach Streuung über Texttypen, wie immer es der Lexikograph für sinnvoll hält). All dies läßt sich automatisch erledigen.

In einem zweiten Schritt wird dieses Vorgehen für die einzelnen Einträge wiederholt: die Belege eines Eintrags werden - diesmal jedoch "von Hand" - unter die entsprechenden bei einem Eintrag vorgesehenen Bedeutungsvarianten (etwa den verschiedenen Lesarten von *Verband*) einsortiert; für bislang nicht vorhandene, aber im Corpus belegte Bedeutungsvarianten muß ein neuer Untereintrag angelegt werden. Letzteres ist grundsätzlich bei den neuen Stichwörtern der Fall. Dieser Bearbeitungsschritt erfordert bereits die lexikalische Kompetenz des Bearbeiters. Er läßt sich nicht vollautomatisieren, sehr wohl allerdings teilautomatisieren und damit im Vergleich zur Arbeit in einem Zettelarchiv erheblich beschleunigen. Dabei mag es sinnvoll sein, zahlreiche Belege, die dieselbe Verwendung bezeugen, gleich "auszusortieren". Das heißt hier jedoch nicht, sie ganz zu entfernen, sondern sie werden einfach als Dopplungen markiert und deshalb im nächsten Schritt nicht berücksichtigt. Man kann aber jederzeit darauf zurückgreifen, beispielsweise wenn man wissen will, wie oft ein Wort in einer bestimmten Verwendung bei welchem Autor oder in welchem Texttyp vorkommt.

Der dritte Schritt gilt nun der eigentlichen Bedeutungsbeschreibung. Dabei kann man sich, wo vorhanden, auf die - in aller Regel sehr klaren - Paraphrasen des WDG stützen, sie gegebenenfalls aufgrund des neuen Belegmaterials oder auch nach Konsultation anderer Wörterbücher modifizieren; für neue Stichwörter bzw. für neue Verwendungen geht dies nicht; es ist eine neue "Definition" erforderlich. Im einen wie im anderen Falle ist die klassische Kompetenz des Lexikographen gefragt; aber auch diese Arbeit läßt sich durch die Teilautomatisierung wesentlich beschleunigen.

Dieses hier in Form dreier Schritte beschriebene Vorgehen veranschaulicht, was oben als "kumulative Entwicklung" bezeichnet wurde. Dabei braucht in keinem dieser Schritte jeweils das gesamte Material des Corpus herangezogen zu werden. Man kann sich durchaus denken, nur Belege ab 1900 heranzuziehen oder auch bestimmte Textsorten zunächst einmal auszuklammern. Dies reduziert fürs erste die Funktionalität. Man kann dann eben nicht nachsehen, was *dalest* bedeutet und wie es Goethe verwendet hat. Aber es ist überschaubar, und es führt für den bearbeiteten Bereich in Abdeckung und Tiefe rasch über all das hinaus, was in konventionellen Wörterbüchern darstellbar und zu finden ist.

10.5 Die semantische Analyse als lexikalisches Lernen

Ein DLS kann die unter 7.1 umrissenenen Probleme nicht schlagartig lösen; aber es kann sie besser lösen als jedes Wörterbuch. Es kann schrittweise auf zunehmende Funktionalität hin ausgebaut werden. Es kann verschiedene Methoden miteinander verbinden. Vor allem aber kann es eines: es kann die Art und Weise, in der wir - vor allem als Kinder, aber auch in späteren Lebensphasen - lexikalisches Wissen erwerben, gezielt nachspielen. Mir scheint überhaupt, die Aufgabe des Lexikographen ist es nicht so sehr zu sagen, was die Bedeutung eines Wortes *ist*, sondern es dem Nutzer möglich zu machen zu verstehen, *was ein Wort zur Gesamtbedeutung eines größeren Ausdrucks beiträgt*. Dieser Beitrag ist nicht konstant: er variiert in der Zeit, er variiert nach Texttyp, und er variiert in der Art und Weise, wie die lexikalische Einheit mit anderen lexikalischen Einheiten der Konstruktion zusammenspielt. In einfachen Fällen läßt sich dieser Beitrag so darstellen, daß der Nutzer ihn leicht erfassen kann; in anderen sind viele variierende Darstellungsformen erforderlich, die erst in ihrer Gesamtheit zu einem Verständnis führen. In jedem Falle ist die Beschreibung nichts Abgeschlossenes, es ist eine Lernhilfe. Dazu ist ein DLS ein ideales Instrument.

11. Schlußbemerkungen

Die Lexikographie ist eine ebenso wichtige wie bewundernswerte Wissenschaft. Sie ist, was die Menge des

akkumulierten Wissens über die Sprachen der Welt, aber auch den praktischen Nutzen für die Vielen angeht, der am weitesten fortgeschrittene Teil der Sprachwissenschaft. Gemessen aber an der Aufgabe, das Lexikon auch nur einer einzigen Sprache in all seinen Aspekten zu beschreiben, ist sie hoffnungslos im Rückstand. Das hat viele Gründe, von denen die Beschränkungen des klassischen Formats der Darstellung, nämlich das gedruckte Wörterbuch, einen wesentlichen Teil ausmachen. Von der Erfindung der Schrift bis in die Gegenwart hat es aber keinen besseren Weg gegeben. Die digitale Technik macht es nun erstmals möglich, die Aufgabe vielleicht nicht zu lösen, aber ihrer Lösung ein guten Schritt näherzukommen. Der Weg dazu führt vom Wörterbuch zum komplexen Digitalen Lexikalischen System, so wie sie hier ecemplarisch beschrieben wurden. Dies ist ein weiter Weg; aber einer der Vorzüge solcher Systeme ist, daß sich die Aufgabe in viele Teilaufgaben zerlegen läßt, die sich getrennt bearbeiten lassen und deren Lösung jeweils ein wichtiger, selbständig nutzbarer Beitrag zum Gesamtziel ist.

Wilhelm von Humboldt schrieb 1810 über das Wesen der wissenschaftlichen Tätigkeit 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.

Summary

From Dictionaries to Digital Lexical Systems

No area in the study of human languages has a longer history and a higher practical significance than lexicography. Ever since its beginnings more than 4500 years ago, this tradition was determined by the format in which the various properties of words can be represented - by the two dimensions of a written or printed document, as best exemplified in the conventional dictionary. In fact, the mere idea that the lexicon of a language can be described in some other form than by a dictionary is difficult to imagine. But the advent of the computer made precisely this possible, in ways which go far beyond the digitization 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 discussed the range of these possibilities and describes the planned German "Digital Lexical System of the Academy", to be realised at the Berlin-Brandenburg Academy of Sciences (www.dwds.de).

Literatur

Adelung, J. Ch. (1777ss): *Versuch eines vollständigen grammatisch-kritischen Wörterbuches der Hochdeutschen Mundart.* Leipzig: Breitkopf.
Bergenholtz, H. und Mugdan, J. (1990): Formen und Probleme der Datenerhebung II: Gegenwartsbezogene synchronische Wörterbücher. In Hausmann u.a. (1989 - 1991), S. 1611 - 1625.
Burger, H. (2003): *Phraseologie. Eine Einführung am Beispiel des Deutschen.* (2. Auflage) Berlin: E. Schmidt.
Clark, E. V. (1993): *The lexicon in acquisition*. Cambridge: Cambridge University Press.
Cowie, Ap. P., Hrsg. (1999): *Phraseology. Theory, Analysis, and Applications.* Clarendon Press: Oxford.
Dietrich, R. (2002): *Psycholinguistik*. Stuttgart: Metzler.
Fellbaum, C. , Hrsg. (1998): *WordNet: an electronic lexical database*. Cambridge, Mass.: MIT Press.
Fillmore, C.J, Johnson, C. R. und Petruck, M.R. L. (2003): Background to FrameNet. In *International Journal of Lexicography 16*, S. 235 - 249.
Hausmann, F-J., Reichmann, O, Wiegand, H. E., Zgusta, L, Hrsg. (1989-1991): *Wörterbücher - Dictionaries - Dictionnaires*. De Gruyter, Berlin, New York.
Horgan, A. D. (1994). *Johnson on Language*. New York: St. Martin's Press.
Hüllen, W. (1999): *English Dictionaries 800 - 1700. The topical tradition.* Oxford: Clarendon Press.
Wilhelm von Humboldt (1906-36): *Gesammelte Schriften.* Berlin: Behr.
Jackson, Howard (2002): *Lexicography. An Introduction.* Routledge, London.
Klein, W. (1974): *Variation in der Sprache*. Kronberg: Scriptor.

Klein, W. (2004): Das digitale Wörterbuch der deutschen Sprache des 20. Jahrhunderts. In J. Scharnhorst, Hrsg., *Sprachkultur und Lexikographie*. Frankfurt/M: Peter Lang, 281 - 309.
Lemberg, I., Schröder, B., und Storrer, A., Hrsg. (2001): *Chancen und Perspektiven computergestützter Lexikographie. Hypertext, Internet und SGML/XML für die Produktion und Publikation digitaler Wörterbücher*. Tübingen: Niemeyer.
Landau, S. A. (2001): *Dictionaries: The Art and Craft of Lexicography*. Scribner, New York.
McArthur, T 1986 *Worlds of Reference. Lexicography, Learning and Language from the clay tablet to the computer.* Cambridge University Press, Cambridge.
Morton, H.C. (1994): *The Story of Webster's Third: Philip Gove's Controversial Dictionary and Its Critics*. Cambridge: Cambridge University Press.
Murphy, M. L. (2003): *Semantic Relations and the Lexicon. Antonymy, Synonymy, and other Paradigms*. Cambridge, Cambridge University Press.
Mugglestone, L., Hrsg. (2000): *Lexicography and the OED*. Oxford, Oxford University Press.
Passow, F. (1831): *Handwörterbuch der griechischen Sprache*. Leipzig.
Pfeifer, W. u.a. (1989): *Etymologisches Wörterbuch des Deutschen*. Berlin: Akademie-Verlag.
Reichmann, O. (1990): Formen und Probleme der Datenerhebung I: Synchronische und diachronische historische Wörterbücher. In Hausmann u.a. (1989 - 1991), S. 1588 - 1611.
Scheller, I. J. G (1783): *Ausführliches und möglichst vollständiges lateinisch-deutsches Lexicon*. Leipzig.
Stevenson, Mark (2003): *Word Sense Disambiguation*. Stanford: CSLI.
Storrer, A. (2001): Digitale Wörterbücher als Hypertexte: Zur Nutzung des Hypertextkonzepts in der Lexikographie. In Lemberg u.a. (2001), S. 88-104.
Ternes, E. (1989): Die phonetischen Angaben im allgemeinen einsprachigen Wörterbuch. In Hausmann u.a. (1989 - 1991), S. 508 - 518.
Wiegand, H.E. (1989): Aspekte der Makrostruktur im allgemeinen einsprachigen Wörterbuch, Der Begriff der Mikrostruktur, Formen von Mikrostrukturen im allgemeinen einsprachigen Wörterbuch (Artikel 38, 38a und 39) in Hausmann u.a. (1989 - 1991), S. 371 - 501.