In: K.-P. Konerding, A, Lehr (Hrsg.): *Linguistische Theorie und lexikographische Praxis. Symposiumsvorträge, Heidelberg 1996*. Tübingen: Max Niemeyer Verlag 1997 (Lexicographica, Series Maior 82).

Peter Hellwig

Ein theorie-übergreifender Standard für lexikalische Wissensbasen

Das zu Ehren von Herbert E. Wiegand abgehaltene Symposium ist der Wechselbeziehung zwischen linguistischer Theoriebildung und praktischer Wörterbucharbeit gewidmet. Mein Fach, die Computerlinguistik, wird von dieser Frage stark tangiert, da die erfolgreiche Lösung einer linguistischen Aufgabe mit dem Computer gleichermaßen vom theoretischen Ansatz wie von der lexikalischen Datenbasis abhängt. Ich denke, es wird daher willkommen sein, wenn ich im ersten Teil meines Vortrags einige Überlegungen über den Zusammenhang von Daten, Theorien und Anwendungen vortrage. Dabei werde ich auch kurz die Frage streifen, welche Anforderungen an Computerlexika gestellt werden, und ob es ganz andere sind als die, die an Wörterbücher für Menschen herangetragen werden.

Als dringendes Problem linguistischer Softwareentwicklung gilt heutzutage die Bereitstellung und die Wiederverwertbarkeit von lexikalischen Ressourcen. Seit die Aussicht besteht, daß die Verarbeitung von Texten auf Computern erhebliche wirtschaftliche Bedeutung erlangt, wird der Ruf nach einer Standardisierung von Computerlexika immer lauter. Im zweiten Teil meines Vortrags werde ich der Frage nachgehen, in wieweit es überhaupt eine Norm für lexikalische Datenbasen geben kann, obwohl die Computerprogramme, die sie benutzen sollen, auf verschiedenen linguistischen Theorien beruhen.

Im dritten Teil des Vortrags will ich dann schon einmal einige Details formulieren, die den zukünftigen Standard charakterisieren werden. Ich stütze mich dabei auf Überlegungen, die in verschiedenen Projekten der EU, in denen ich mitgearbeitet habe, angestellt worden sind.

1. Daten, Theorie und Anwendung

> Nothing is so practical as a good theory, and there are few better theoretical challenges than a precise practical problem, empirically unsolvable by mere practicalities.
>
> Hans Karlgren

Ich möchte meine Erörterung des Verhältnisses von Daten, Theorie und Anwendung mit einer Anekdote beginnen. Während einer Tagung in Pisa, bei der es um gemeinsame Lexika für unterschiedliche Grammatiken ging, folgten wir einer Mitarbeiterin auf dem Weg vom Hotel zum Institut. Ich versuchte mir jede einzelne Kreuzung einzuprägen, an der unsere Führerin wieder einmal die Richtung wechselte, damit ich später den Weg allein zurück finden könnte. Bald war ich jedoch geneigt, vor der "Datenflut" der verwinkelten Gassen zu kapitulieren. Da äußerte einer der mitmarschierenden Kollegen eine Hypothese: Die Mitarbeiterin schlage in den rechtwinklig aufeinanderstoßenden Gassen einen Weg ein, der möglichst nahe zu einer gedachten Diagonale durch die Häuserblocks verliefe, und sie wechsele deshalb so gut wie an jeder Kreuzung die Richtung. Indem ich auf dem Heimweg diese "Diagonalen-Theorie" anwandte, erreichte ich - mit einer nur geringen Abweichung - allein wieder das Hotel.

Was lehrt diese Episode? Im Leben ist die Fülle von Daten überwältigend. Das Wesen einer Theorie liegt darin, in Bezug auf eine bestimmte Fragestellung aus der Fülle von Daten ein Prinzip (oder eine Menge von Prinzipien) zu abstrahieren. Aufgrund von Prinzipien lassen sich Vorhersagen über die Realität machen, und indem Vorhersagen gemacht werden können, lassen sich praktische Probleme meistern. Legt man ein solches Verständnis von Theorie zugrunde und versteht unter Praxis das Lösen von Aufgaben im wirklichen Leben, ist es unmöglich, Theorie und Praxis gegeneinander auszuspielen. Wenn man es doch tut, meint man, daß jemand in seiner Praxis einer anderen, eventuell uneingestandenen Theorie folgt, als der vorgeblichen. Als Vorhersage unter einer bestimmten Fragestellung ist aber jede Theorie genuin mit einer Praxis verknüpft, und jede praktische Handlung, soweit sie Prinzipien folgt, impliziert per definitionem eine Theorie.

Theorie und Praxis sind also das eine, die realen Daten sind das andere. Was natürlich passieren kann, ist, daß Theorie und Praxis nicht mit der realen "Datenbasis" übereinstimmen. Zwischen Realität und Theorie ebenso wie zwischen Realität und praktischem Handeln kann es natürlich Konflikte geben.

Zumindest für die Computerlinguistik ist die geschilderte wissenschaftstheoretische Position kennzeichnend. Letztlich sind ausführbare Computerprogramme das Ziel, d.h. Lösungen für praktische Aufgaben, sei es eine akkurate Silbentrennung, die Prüfung eines Textes auf Grammatikfehler, die Suche nach einer Information, eine Übersetzung. Das jeweilige Computerprogramm ist eine in einen Algorithmus gegossene linguistische Theorie, die bezüglich der sprachlichen Ein- und Ausgabedaten die erwünschten Vorhersagen macht.

Während die enge Verbindung von Theorie und Aufgabenstellung m.E. allgemein gilt, kommt bei der Computerimplementierung noch hinzu, daß es sich um eine formalisierte Theorie handeln muß, d.h. Vorhersagen werden "generiert", indem Symbole nach feststehender Vorschrift manipuliert (erzeugt, verglichen, ersetzt, gelöscht) werden. Die Menge der Symbole, ihre Kombination und die Vorschriften zu ihrer Manipulation ergeben den sogenannten Formalismus. Da eine natürliche Sprache auf der Oberfläche ja ebenfalls aus Symbolen besteht, ist es möglich, die Menge dieser Symbole mit in den Formalismus aufzunehmen und an bestimmten Schnittstellen nur Folgen von natürlich-sprachlichen Symbolen zuzulassen - fertig ist das sprachverarbeitende System! Während die eigentlichen Operationen innerhalb des Programms in einer Programmiersprache beschrieben werden, ist es üblich, die natürlich-sprachlichen Symbole sowie ihre meta-sprachlichen Beschreibungen in Grammatiken und Lexika zu speichern, die vom Programm bei Bedarf eingelesen werden.

Aus dieser Sachlage ergibt sich folgendes: Ein Computerlexikon, das unmittelbar an ein Anwendungsprogramm angeschlossen ist, muß außerordentlich eng mit dem Algorithmus für diese Anwendung verzahnt sein. Es muß ja genau die Symbole bereitstellen, die nach der zugrundeliegenden Theorie möglichst effizient manipuliert werden können, um die gestellte Verarbeitungsaufgabe zu lösen. Auf dieser Ebene des internen Gebrauchs sind Computerlexika nicht nur ziemlich verschieden von Wörterbüchern, die ein Mensch benutzt, sie unterscheiden sind auch untereinander, je nach der Aufgabe, die der Algorithmus lösen soll, und je nach den theoretischen Prinzipien, die dabei zugrunde gelegt werden.

2. Ein polytheoretisches Lexikon?

Angesichts der Theorie- und Aufgabenabhängigkeit der Computerlexika ist es nicht verwunderlich, daß es bisher noch kaum Austausch von lexikalischen Ressourcen zwischen verschiedenen Computeranwendungen gibt. Meist wird für eine neue Anwendung erst einmal ein neues Computerlexikon gemacht. Viele solcher Lexika erreichen erst gar nicht den notwendigen Umfang für einen realistischen Einsatz in der Praxis. Das ist mit ein Grund, daß bisher erst ganz wenige sprachverarbeitende Programme Produktreife erlangt haben und im Alltag eingesetzt werden.

Dies ist natürlich ein unbefriedigender Zustand. Die *Language Industry* könnte so schöne Zuwachsraten haben, *Language Engineering* macht's möglich, wenn da nicht der Engpaß der lexikalischen Informationen über Hunderttausende von Wörtern in etlichen Sprachen wäre. Der

Computerlexikographie selbst könnte eine rosige wirtschaftliche Zukunft beschieden sein, wenn ihre Erzeugnisse in vielen praktisch eingesetzten Programmen benutzt würden. Damit aber Angebot und Nachfrage von lexikographischen Daten in der elektronischen Welt miteinander kompatibel sind, bedarf es dringend einer Norm.

Seit ungefähr zehn Jahren fördert die Europäische Gemeinschaft Projekte, die die Voraussetzungen für den Austausch lexikalischer Ressourcen untersuchen und Empfehlungen für ihre Standardisierung ausarbeiten sollten. Zu nennen ist u.a. die großangelegte Studie EUROTRA-7 ("Feasibility Study on the Reusability of Lexical and Terminological Resources in Computerized Applications"), an der zahlreiche Universitätsinstitute beteiligt waren. Leider existieren die Ergebnisse nur in Form von "grauer" Literatur. Dazu zählte weiter das Esprit-Projekt AQUILEX, das den Import von Daten aus gedruckten Wörterbüchern zum Ziel hatte, außerdem MULTILEX, ein Projekt, das ein Lexikonformat für mehrsprachige Anwendungen und dazu passende Werkzeuge zur Lexikonpflege entwickeln sollte, und vor allem das Eureka-Projekt GENELEX, in dem ein detailliertes Modell für linguistische Datenbanken entworfen wurde. In den vergangenen drei Jahren versuchte die Arbeitsgruppe EAGLES (Expert Advisory Group on Language Engineering Standards) die Ergebnisse all dieser Projekte zu bündeln. Obwohl es schwer zu durchschauende Rivalitäten gibt, sieht es zur Zeit so aus, als ob die GENELEX-Vorschläge im Kern am meisten Aussicht haben, zu einem Standard für elektronische Lexika zu werden. Darin enthalten sind auch konkrete Vorschriften für ein Austauschformat, welches die Markierungssprache SGML benutzt und mehr oder weniger in Übereinstimmung mit der weltweiten Text Encoding Initiative (TEI) steht. Die TEI arbeitet auch auf eine einheitliche Kennzeichnung linguistischer Korpora hin. Das neueste Projekt der EU ist PAROLE (Preparatory Actions for Linguistic Resources Organization for Language Engineering). Im Rahmen von PAROLE sollen nun nationale Gruppen die Richtlinien von EAGLES und GENELEX auf die jeweilige Sprache anwenden. Für das Deutsche ist das Institut für deutsche Sprache in Mannheim bei dieser Aktion federführend.

Die Details, die ich im dritten Teil meines Vortrags vorstellen will, entsprechen dem gegenwärtigen Hauptstrom in den EU-Projekten. Zunächst muß ich aber einige Stationen auf dem Weg zum gegenwärtigen Stand nachzeichnen. Die Standardisierung von Computerlexika ist nämlich anders verlaufen, als gedacht, und das Ergebnis wird in der Computerlinguistik auch noch keineswegs von allen akzeptiert. Da es dabei vor allem um linguistische Theoriebildung und praktische Lexikographie geht, hat die Geschichte vielleicht auch für dieses Symposium einen Erkenntniswert.

Anfang der achtziger Jahre war, nach langem Vorherrschen der generativen Transformationsgrammatik, eine Bewegung in die Grammatikformalismen gekommen, die man als Konvergenz interpretieren konnte. Nicht zuletzt unter dem Einfluß der Computerimplementierung entstanden formale Grammatiken, die sich in vielem ähnelten, obwohl sie auf unterschiedliche Traditionen zurückgingen, z.B. auf die amerikanische Phrasenstruktursicht und den eher europäischen Dependenzansatz. Schon sprach man z.B. von einer "Familie" der Unifikationsgrammatiken. Kennzeichnend war für die Entwicklung allgemein eine lexikalistische Sicht, d.h. ein Großteil der grammatischen Information wird seither im Lexikon untergebracht. Das Lexikon enthält nun nicht mehr die Ausnahmen von der Regel sondern die Regeln selbst. Wenn es in der Grammatik überhaupt noch eine eigene Regelkomponente gibt, so ist sie sehr abstrakt.

Unter dem Eindruck dieser Konvergenz der Formalismen entstand die Idee eines "polytheoretischen" Lexikons, das für alle Projekte die notwendige lexikalische Information zur Verfügung stellen sollte. Nicht theorie-unabhängig sollte das Lexikon sein, sondern eben poly-theoretisch, das heißt allen theoretischen Anforderungen genügend. Es gab mehrere Konferenzen, an denen Vertreter verschiedener grammatischer Schulen teilnahmen. Schließlich mündeten die Bemühungen ein in das schon genannte EUROTRA-7 Projekt.

Die ursprünglich eingeschlagene Vorgehensweise könnte man als den "Normalform"-Ansatz charakterisieren. Ziel war sozusagen ein Super-Formalismus, der schwach-äquivalent zu allen beteiligten Einzelformalismen sein sollte, d.h. in unserer obigen Redeweise: Er sollte genau dieselben Vorhersagen erlauben wie jede einzelne Grammatiktheorie. Außerdem sollte es ein formales Verfahren geben, mit dem die Aussagen jeder Einzeltheorie in den Super-Formalismus überführt werden können. In dieser Normalform sollte auch das poly-theoretische Lexikon geschrieben werden. Indem exakte Umformungsregeln zwischen Einzelformalismus und Normalform aufgestellt würden, könnte jeder dieselbe lexikalische Basis für seine speziellen Anwendungen benutzen.

Als erster Schritt, um zu einer Normalform zu kommen, wurden in der EUROTRA-7 Studie alle Kategorien verglichen, welche in ausgewählten formalen Grammatiktheorien (u.a. GB, GPSG, HPSG, DUG, EUROTRA) vorkamen. Wenn wir uns den ersten Teil meines Vortrags ins Gedächtnis rufen, sind wir nicht sehr überrascht, daß das Vorhaben bereits hier scheiterte.

Die oben geschilderte Episode auf dem Weg vom Hotel zur Tagungsstätte in Pisa hätte auch so enden können: "Da äußerte einer der mitmarschierenden Kollegen eine Hypothese: Die Mitarbeiterin schlage in

den rechtwinklig aufeinanderstoßenden Gassen einen Weg ein, der möglichst nahe *zu einer Linie verliefe, die der Himmelsrichtung Süd-Süd-West entspräche*, und sie wechsele deshalb so gut wie an jeder Kreuzung die Richtung. Indem ich auf dem Heimweg diese "*Süd-Süd-West*-Theorie" anwandte, erreichte ich - mit einer nur geringen Abweichung - allein wieder das Hotel."

Die Diagonalentheorie und die Himmelsrichtungstheorie sind anscheinend schwach-äquivalent, denn sie führen zum gleichen Ergebnis. Die Prinzipien beider Theorien beruhen aber auf völlig verschiedenen Generalisierungen, nämlich einmal auf dem Verlauf des Weges relativ zu den quadratischen Häuserblocks, das andere Mal relativ zum Nordpol und den Gestirnen. Die eine Theorie ist nicht "wahrer" als die andere. Man könnte einwenden, die Himmelrichtungstheorie sei überlegen, da sie allgemeiner anwendbar sei. Das macht die Diagonalentheorie aber nicht überflüssig. Sofern ich keinen Kompaß habe und die Sonne nicht scheint, werde ich mich besser an der gedachten Diagonale durch die Häuserblocks orientieren. Es dürfte schwer fallen, ein Prinzip aufzustellen, das als "Normalform" für beide Theorien gelten könnte. Das Gemeinsame beider Theorien ist nur das Ergebnis. Wie findet man heraus, ob zwei Theorien zum selben Ergebnis führen? Indem man sie auf reale Daten anwendet!

Dasselbe wurde deutlich beim Vergleich verschiedener Formalismen im Rahmen der EUROTRA-7 Studie. Formale Theorien sind deduktive Systeme; jede Kategorie ist nur durch die Gesamtheit der Ableitungen definiert, in die die Kategorie eingeht. Die Ableitungen (oben habe ich sie "Symbolmanipulationen" genannt) realisieren die jeweiligen Generalisierungen der Theorie. Wie vergleicht man unter diesen Umständen die Kategorien zweier Formalismen? In der Praxis gab es nur einen Weg: Suche ein Beispiel, auf das die Kategorie der einen Theorie zutrifft. Suche ein Beispiel, auf das die Kategorie der anderen Theorie zutrifft. Vergleiche die Beispiele.

Die exemplarische Extension der Kategorien einiger gängiger Grammatikformalismen ergab bald, daß die Ableitungsmechanismen und die Generalisierungen viel zu verschieden sind, um daraus eine Normalform zu konstruieren, in der ein polytheoretisches Lexikon abgefaßt werden könnte. Gleichzeitig ergab sich aber auch eine neue Perspektive. Zum Zwecke des Vergleichs der Kategorien war man gezwungen gewesen, auf die Beobachtungsebene hinabzusteigen. Man hatte den Weg der Generalisierung umgekehrt beschreiten müssen, von den mehr oder weniger abstrakten Kategorien der Theorien hin zu den Daten, die sie abdeckten. Bei der Überprüfung der empirischen Daten wandte man dann unwillkürlich Kriterien der Opposition und Distribution an, Entscheidungsprozeduren wie Kommutation, Permutation, Para-

phrasetests - kurz das Instrumentarium des taxonomischen Strukturalismus. Die observierende Beschreibung der Distribution sprachlicher Phänomene ist offensichtlich eine Ebene, auf der auch Linguisten verschiedener Provenienz meist eine Verständigung erreichen. Sollte man diese Ebene dann nicht gleich zur angestrebten Schnittstelle zwischen den Theorien machen? Das polytheoretische Lexikon wäre dann nicht als Überbau sondern als Unterbau für alle Theorien zu konzipieren, eine Daten-"Basis" im wörtlichen Sinne.

Ich erhebe diesen Gedankengang zur These: Bei den lexikalische Ressourcen, die zwischen verschiedenen Anwendungen austauschbar sind und die deshalb Gegenstand von Standardisierungsbemühungen sein sollten, kann es sich nur um Beobachtungsdaten handeln. Um zu diesen Beobachtungsdaten zu kommen, ist auf die Methoden des taxonomischen Strukturalismus zurückzugreifen.

Diese These ist provozierend. Gilt der Deskriptivismus nicht als Irrtum? Wurde der taxonomische Strukturalismus nicht vor langem von der generativen Grammatik überwunden? Ist es somit nicht ein Rückschritt, das polytheoretische Lexikon auf diese Position zu gründen? Die Anstrengungen in der Computerlinguistik gehen gerade in die umgekehrte Richtung, nach mehr Generalisierung, allgemeineren Prinzipien, Unifikation, Vererbung, Beseitigung von Redundanz.

Die Kritik am taxonomischen Strukturalismus wäre berechtigt, wenn es um die Theorien und Anwendungen selbst ginge. Die sollen aber durchaus weiter so bleiben wie bisher. Jede einzelne Schule soll aus der Datenbasis der empirischen Befunde die theoretischen Prinzipien ableiten, die sie zu erkennen glaubt, und die für die Lösung der gestellte Aufgabe hinreichend sind. Dabei wird jede Anwendung aus der allgemeinen Wissensbasis ihr eigenes Speziallexikon erzeugen, das - wie bisher- genau auf den jeweiligen Algorithmus abgestimmt ist. Übrigens wäre die formale Ableitung der theoretischen Konstrukte aus explizit kodierten Beobachtungsdaten eine heilsame Übung. Es ist zu vermuten, daß manche Theorie in sich zusammenfallen würde, wenn sie ihre empirische Basis ausbuchstabieren müßte.

Der Gedanke an einen rein deskriptiven lexikalischen Standard ist in der Computerlinguistik noch gewöhnungsbedürftig. Wer aber im Gegensatz dazu vorschlägt, europäische Sprachressourcen auf der Ebene der Formalismen zu standardisieren, strebt in Wirklichkeit, bewußt oder unbewußt, die Hegenomie einer einzigen theoretischen Schule an. Die Konkurrenz wird aber nicht zulassen, daß eine Theorie bereits das Vorfeld usurpiert, d.h. ein anerkannter Standard für eine umfassende lexikalische Wissensbasis wird sich einfach nicht etablieren lassen, wenn er bereits theoretische Vorentscheidungen einschließt.

Ich stelle zur Diskussion, ob man meine auf die Computerlinguistik zugeschnittene These verallgemeinern kann. Das Symposium ist der Wechselwirkung zwischen linguistischer Theoriebildung und praktischer Wörterbucharbeit gewidmet. Sollte die allgemeine Lexikographie im Sinne meiner These ebenfalls nur die Beobachtungsdaten liefern? In der Praxis ist es ja wohl sowieso so. Der Theoretiker sollte die Lexikographen nicht deswegen schelten. Zu fordern ist nur, daß die gesammelten Daten eine verläßliche Arbeitsgrundlage für die Theoriebildung darstellen. Wie die Datenerhebung methodisch so abgesichert werden kann, daß diese Forderung eingelöst wird, ist eine Frage, über die noch viel nachzudenken und zu sagen ist. Die Meta-Lexikographie behält durchaus ihren Gegenstand.

Hier erwarte ich natürlich den Einwand, daß jede Beobachtung von einem bestimmten Interesse ausgeht und von Vorannahmen beeinflußt wird, daß der Datenerhebung ohne theoretischen Hintergrund der begriffliche Apparat fehlt. Diese Kritik ist ja gegen den taxonomischen Strukturalismus vorgebracht worden und ist auch völlig berechtigt. Die These, daß sich der Lexikograph um die Daten kümmern soll, während der Theoretiker sich mit Problemlösungen beschäftigt, darf nicht so mißverstanden werden, daß sich die lexikographische Praxis von der linguistische Theoriebildung abkoppeln sollte, ganz im Gegenteil. Es bedarf der stetigen Kommunikation, damit die Lexikographen wissen, welche Beoachtungsdaten die Theoretiker brauchen.

Um es im Bild meines Weges durch Pisa auszudrücken: Als Datenbasis für die "Diagonaltheorie" ist die Form der Häuserblocks relevant, während die "Süd-Süd-West-Theorie" den Winkel zwischen den Wegen und der Richtung zum Nordpol untersuchen muß. Wenn eine der Informationen nicht zur Verfügung stünde, wäre die Datenbasis zu schmal, um beiden Theorien zu dienen. In Wirklichkeit ist es allerdings so, daß beides in einem vernünftigen Stadtplan von Pisa schon enthalten ist.

Kriterien für die Datenerhebung müssen also von der Theorie kommen. Dies muß geschehen, ohne die Lexikographie für eine bestimmte Theorie zu vereinnahmen. Es m.E. eine Bringschuld der jeweiligen Theorie, sich sozusagen "in die Niederungen ihrer eigenen empirischen Voraussetzungen" hinabzubegeben und objektive (d.h. inter-subjektiv nachvollziehbare) Entdeckungsprozeduren anzugeben, die der Lexiko-graph einsetzen kann.

3. Datenbankmodell und Austauschformat

Wie soll der Standard für wiederverwertbare lexikalische Ressourcen für die maschinelle Sprachverarbeitung nun konkret aussehen? Zum

gegenwärtigen Zeitpunkt kann man m.E. schon zweierlei sagen: Konzeptionell wird es sich um ein Datenbankmodell handeln, und physikalisch wird der Standard als Austauschformat für Dokumente in der Sprache SGML realisiert werden.

Datenbanken bestehen im Prinzip aus Tabellen, die hierarchisch untergliedert und/oder über Querverweise zwischen ihren Einträgen miteinander verbunden sein können. Tabellen bieten sich für einfache und uniforme Massendaten an. Auch Wörterbücher sind normalerweise nach diesem Prinzip organisiert. Datenbanken sind aber keineswegs die bevorzugten Wissensrepräsentation in der Computerlinguistik und der linguistischen Informatik. Moderne "wissensbasierte" Systeme enthalten viel eher Formeln eines Logikkalküls, die von einem Theorembeweiser ausgewertet werden, oder netzwerkartige Graphen mit Knoten und Kanten, die einer Mustererkennung unterzogen werden. Ich wiederhole mich, wenn ich sage: Das soll auch so bleiben. Diese Datenstrukturen sind wahrscheinlich für die jeweilige Aufgabenstellung optimal. Nur eignen sie sich schlecht für den Austausch von lexikalischen Ressourcen.

Ich plädiere für ein Datenbankmodell als Standard aus zwei Gründen. Praktisch spricht dafür, daß sehr große Mengen von Daten anfallen werden, zu deren Verwaltung stabile, kommerzielle Software zur Verfügung stehen sollte. Theoretisch spricht dafür, daß relationale Datenbanken so gestaltet werden können, daß sie eine Vielzahl von "Sichten" auf die Daten erlauben und daher wenig über deren Benutzung präjudizieren. Wir wollen ja die lexikalische Datenbasis von theoretischen Vorannahmen möglichst freihalten, weil nur so eine allgemeine Akzeptanz zu erwarten ist.

Eine Datenbank konzipiert man zunächst unabhängig von der Implementierung. Dabei ist die Erstellung eines Entity-Relationship Modell eine gängige Methode. (Es gibt dazu leicht erreichbare Literatur.) Ein solches Modell besteht u.a. aus folgenden Konstrukten:

- *Entitäten,* das sind Objekte, die in der Beschreibung unterschieden werden sollen;

- *Attribute*, das sind Typen von Merkmalen, die Objekte charakterisieren;

- *Attributwerte*, das sind konkrete Merkmale, die einem Objekt zukommen;

- *Relationen,* das sind Beziehungen zwischen Objekten.

Hilfreich ist es, das konzeptuelle Modell graphisch zu skizzieren. Abbildung 1 zeigt zum Beispiel das Schema für ein sehr einfaches Lexikon. Es enthält die Entitäten *Lexikalischer Eintrag, Morphologische Beschreibung,*

Bedeutungsbeschreibung, Wortform, Wortart und Grammatisches Merkmal mit ihren Attributen. Die morphologische Beschreibung besteht hier aus einer einfachen Auflistung der Wortformen. Zu jeder Wortform kann die Wortart und zu dieser eine Menge grammatischer Merkmale angegeben werden.

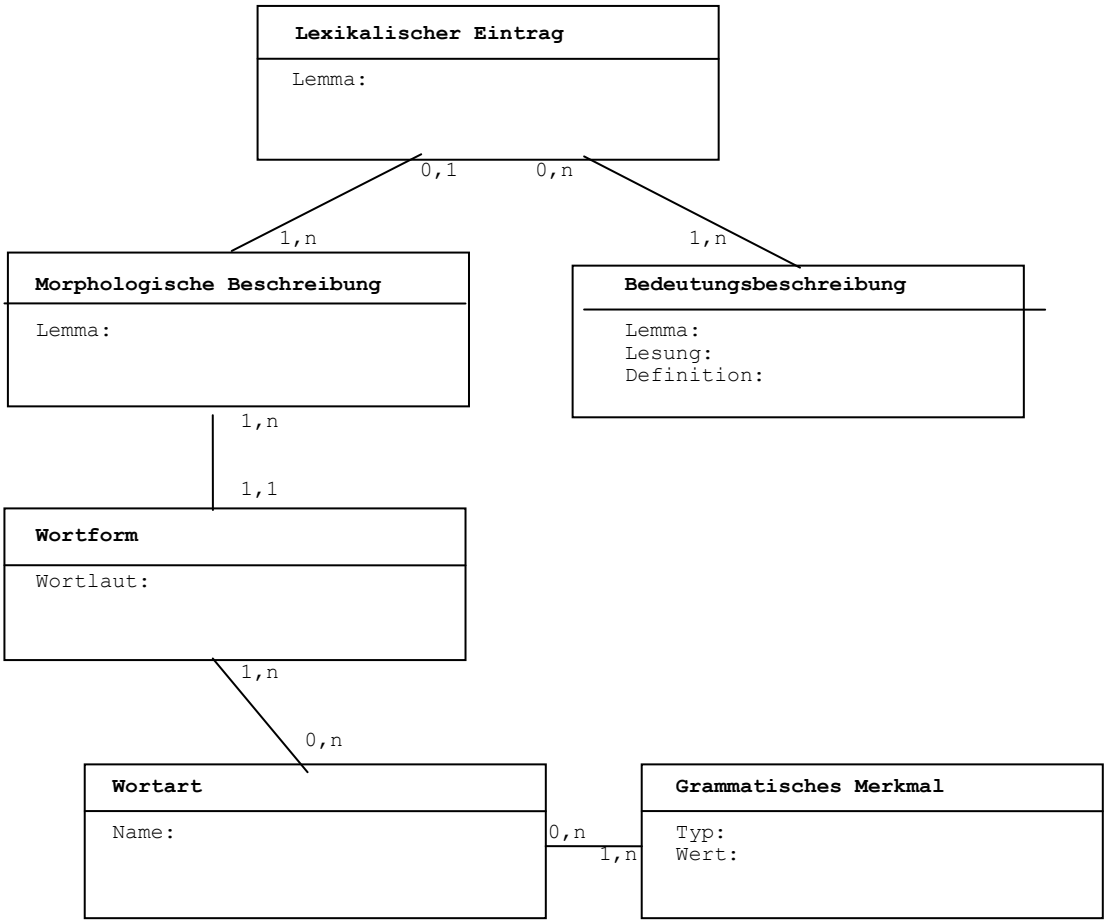

Abbildung 1. Schema eines einfachen Wortformenlexikons

In der gedruckten Form meines Vortrags sind anspruchsvollere Beispiele nicht möglich, weil sie längerer Erklärung bedürften. Ich will hier nur das allgemeine Vorgehen schildern. Zunächst mag jeder, der bereits umfangreiche Computerlexika besitzt, sich klar werden, in welche Entitäten und Attribute die Daten zerlegt werden können und welche Entitäten paarweise zusammenhängen. Im Projekt GENELEX sind Lexika aus verschiedenen Anwendungen und zu verschiedenen Sprachen erfolgreich in derartige Datenbankschemata umgeformt worden. Manchmal muß man sich erst etwas daran gewöhnen, Strukturgebilde oder gar operationale Beschreibungen, wie z.B. Ersetzungsregeln, in eine rein taxonomische Sicht zu überführen, d.h. daran, den Objektbereich nur

mithilfe von Identifikation und Klassifikation von Entitäten darzustellen. Aber da eine Struktur eine Menge von Elementen und Relationen ist und da Relationen als Mengen von Paaren aufgefaßt werden können, sollte die Übersetzung in der Regel gelingen. Wichtig ist, daß man gleichzeitig die heuristischen Kriterien notiert, die zur Identifikation einer jeden Entität und eines jeden Attributwertes hinreichen. Diese Kriterien finden sich später wieder im Instruktionsbuch für die Datenaquisition.

Es ist Aufgabe der Experten, sich über die Entitäten, Attribute und Relationen klar zu werden, die zum Standard gehören sollen. Viele Einteilungen (Morph, Morphem, Lexem, Stamm, Präfix, Infix, Suffix, Wortart, Valenzangabe usw.) kann man aus dem Grundkurs Linguistik übernehmen. Andere Entitäten mögen aus der Perspektive einzelner Anwendungen und Theorien zusätzlich vorgeschlagen werden. Unterschiedliche Einteilungen, soweit es heuristische Kriterien dafür gibt, sind nicht so schlimm. Ja, der Standard sollte durchaus alternativ zu benutzenden Entitäten und Attribute bereitstellen. Ein typischer Unterschied zwischen existierenden Ressourcen kann z.B. in einer gröberen oder feineren Beschreibung liegen. Der Standard muß hier alle Grade zulassen.

Der Austausch standardisierter lexikalischer Daten soll möglichst einfach und unabhängig von bestimmter Software sein. In der Regel sollten die Daten als ASCII-File übergeben werden. Allgemein ist es üblich, in ASCII-Dokumenten besondere Fonts, Zeichen, Formate und auch inhaltlich definierte Einheiten, wie Titel, Kapitel, Überschriften, Zusammenfassung, Register usw., dadurch kenntlich zu machen, daß explizite Angaben dazu (sogenannten "Tags") in den Text eingefügt werden. Die Markierungssprache, die in jüngster Zeit einen Siegeszug angetreten hat, ist SGML. Mit SGML-Tags annotierte ASCII-Texte sind denn auch das Austauschformat, das für lexikalischen Wissenbasen voraussichtlich zur Norm werden wird.

Eine wichtige Regel ist es, daß jeder SGML-markierte Text am Anfang eine Definition aller im Text verwendeter Tags enthalten muß. Dieser Vorspann wird DTD (Document Type Definition) genannt und ist mit der Deklaration der Variablen in einem Computerprogramm vergleichbar. Jeder Text führt so seine eigene SGML-Syntaxbeschreibung mit sich. Erschließungssoftware konsultiert zunächst den DTD und kann so beliebig strukturierte Texte verarbeiten.

Mithilfe geeigneter Tags kann man natürlich auch beliebige inhaltliche Unterscheidungen machen. Man kann daher auch das konzeptuelle Modell der lexikalischen Datenbank in einen SGML-Text übersetzen und später einen solchen SGML-Text in eine aktuelle Datenbank laden. Ein DTD erlaubt dieselben Festlegungen wie eine kontextfreie Grammatik. Somit

ist er nicht nur dazu geeignet, die benutzten Markierungen zu erklären, sondern auch dazu, das zunächst nur konzeptuell und graphisch entworfene Modell der Datenbank exakt zu formalisieren.

Die in SGML bereitgestellten Mittel zur Markierung von Texten passen gut zur Entity-Relationship Sicht einer Datenbank. Entitäten (in der SGML-Terminologie heißen sie "Elemente") werden durch Tags markiert. In der Regel besteht ein Tag aus dem Namen der Entität und ist in spitze Klammern eingeschlossen. Das Ende der Entität wird u.U. durch einen weiteren Tag markiert, der aber zusätzlich einen Schrägstrich enthält. Attribute und Attributwerte der Entität werden einfach in den Tag mit hineingenommen. Für die Darstellung von Relationen zwischen Entitäten gibt es zwei Möglichkeiten. Hierarchische Verhältnisse werden direkt durch Einbettung entsprechender Tags repräsentiert. Verweise kann man dadurch realisieren, daß die Werte bestimmter Attribute in den entsprechenden Entitäten identisch sind. Solche Attribute werden in einer relationalen Datenbank dann als Schlüssel verwendet, mit deren Hilfe man die Verbindung herstellen kann.

Im DTD werden all diese Konstrukte formal festgelegt. Ich erspare mir hier, einen solchen DTD vorzustellen, da ich sonst weitere Details erklären müßte, die für meinen Argumentationszusammenhang unwichtig sind. Abbildung 2 zeigt statt dessen beispielhaft, wie ein Eintrag nach dem konzeptuellen Modell von Abbildung 1 für das Lemma *Hahn* in SGML aussehen könnte:

```
<Lexikalische_Einheit Lemma="Hahn">

<Bedeutungsbeschreibung Lemma="Hahn" Lesung=1>das
maennliche Tier der Haushuehner</Bedeutungsbeschreibung>

<Bedeutungsbeschreibung Lemma="Hahn" Lesung=2>Vorrichtung
zum Sperren und Oeffnen von
Rohrleitungen</Bedeutungsbeschreibung>

<Morphologische_Beschreibung Lemma="Hahn">
     <Wortform> 'Hahn'
          <Wortart Name=Substantiv>
               <Grammatisches_Merkmal Typ=Genus
               Wert=Maskulinum>
               <Grammatisches_Merkmal Typ=Numerus
               Wert=Singular>
               <Grammatisches_Merkmal Typ=Kasus
               Wert=Nominativ>
          </Wortart>
```

```
      <Wortart Name=Substantiv>
            <Grammatisches_Merkmal Typ=Genus
            Wert=Maskulinum>
            <Grammatisches_Merkmal Typ=Numerus
            Wert=Singular>
            <Grammatisches_Merkmal Typ=Kasus
            Wert=Dativ>
      </Wortart>
      <Wortart Name=Substantiv>
            <Grammatisches_Merkmal Typ=Genus
            Wert=Maskulinum>
            <Grammatisches_Merkmal Typ=Numerus
            Wert=Singular>
            <Grammatisches_Merkmal Typ=Kasus
            Wert=Akkusativ>
      </Wortart>
</Wortform>
<Wortform> 'Hahns'
      <Wortart Name=Substantiv>
            <Grammatisches_Merkmal Typ=Genus
            Wert=Maskulinum>
            <Grammatisches_Merkmal Typ=Numerus
            Wert=Singular>
            <Grammatisches_Merkmal Typ=Kasus
            Wert=Genitiv>
      </Wortart>
</Wortform>
<Wortform> 'Hahnes'
      <Wortart Name=Substantiv>
            <Grammatisches_Merkmal Typ=Genus
            Wert=Maskulinum>
            <Grammatisches_Merkmal Typ=Numerus
            Wert=Singular>
            <Grammatisches_Merkmal Typ=Kasus
            Wert=Genitiv>
      </Wortart>
</Wortform>
<Wortform> 'Hahne'
      <Wortart Name=Substantiv>
            <Grammatisches_Merkmal Typ=Genus
            Wert=Maskulinum>
            <Grammatisches_Merkmal Typ=Numerus
            Wert=Singular>
            <Grammatisches_Merkmal Typ=Kasus
            Wert=Dativ>
      </Wortart>
</Wortform>
```

```
<Wortform> 'Haehne'
      <Wortart Name=Substantiv>
            <Grammatisches_Merkmal Typ=Numerus
            Wert=Plural>
            <Grammatisches_Merkmal Typ=Kasus
            Wert=Nominativ
      </Wortart>
      <Wortart Name=Substantiv>
            <Grammatisches_Merkmal Typ=Numerus
            Wert=Plural>
            <Grammatisches_Merkmal Typ=Kasus
            Wert=Genitiv>
      </Wortart>
      <Wortart Name=Substantiv>
            <Grammatisches_Merkmal Typ=Numerus
            Wert=Plural>
            <Grammatisches_Merkmal Typ=Kasus
            Wert=Akkusativ>
      </Wortart>
</Wortform>
<Wortform> 'Haehnen'
      <Wortart Name=Substantiv>
            <Grammatisches_Merkmal Typ=Numerus
            Wert=Plural>
            <Grammatisches_Merkmal Typ=Kasus
            Wert=Dativ>
      </Wortart>
</Wortform>
</Morphologische_Beschreibung>
```

Abbildung 2: Ein Eintrag für das Lemma *Hahn* in SGML-Format

Schon dadurch, daß sich jeder an SGML als Markierungssystem hält, ist gewährleistet, daß andere seine Daten leicht weiterverarbeiten können. Dazu gibt es Software-Werkzeuge, übrigens auch solche, welche die Konsistenz der Lexikonkodierung mit dem einmal aufgestellten DTD prüfen. Noch besser wäre es, wenn man sich für lexikalische Daten auf einen, möglicherweise recht umfangreichen und flexiblen, DTD einigte, in dem auch die Namen der Entitäten und Attribute genormt sind. Dann könnte man nämlich Software entwickeln, die unmittelbar auf die Inhalte der ausgetauschten Daten Bezug nehmen kann, weil die Tags, die diese Inhalte markieren, bekannt sind. Dies genau ist das Ziel gegenwärtiger Bemühungen in der europäischen Projektlandschaft.