LEXICOGRAPHICA Series Maior

LEXICOGRAPHICA

Series Maior

Supplementary Volumes to the International Annual for Lexicography
Suppléments à la Revue Internationale de Lexicographie
Supplementbände zum Internationalen Jahrbuch für Lexikographie



45

Published in cooperation with the Dictionary Society of North America
(DSNA) and the European Association for Lexicography (EURALEX)

Matthias Heyn

# Zur Wiederverwendung maschinenlesbarer Wörterbücher

Eine computergestützte metalexikographische Studie am Beispiel der elektronischen Edition des »Oxford Advanced Learner's Dictionary of Current English«

Max Niemeyer Verlag
Tübingen 1992

*Für Dagmar*

Die Deutsche Bibliothek – CIP-Einheitsaufnahme

*Heyn, Matthias :* Zur Wiederverwendung maschinenlesbarer Wörterbücher : eine computergestützte metalexikographische Studie am Beispiel der elektronischen Edition des »Oxford advanced learner's dictionary of current English« / Matthias Heyn. – Tübingen : Niemeyer, 1992
(Lexicographica : Series maior ; 45)
NE: Lexicographica / Series maior

ISBN 3-484-30945-8 ISSN 0175-9264


Druck: Weihert-Druck GmbH, Darmstadt
Einband: Hugo Nädele, Nehren

# Inhaltsverzeichnis

# Vorwort

Dieses Buch wendet sich vor allem an Linguisten, Lexikologen, Lexikographen und Informatiker, die mit Interesse die Diskussion um die Wiederverwendung sprachlicher Informationen aus maschinenlesbar vorliegenden Wörterbüchern verfolgen. Diese Diskussion wird vorwiegend in dem jungen und sich erst formierenden Fachbereich der Computerlexikographie geführt und befaßt sich mit der Problematik, ob und wie die Informationen, die in ein- oder mehrsprachigen Wörterbüchern vorgefunden werden, im Kontext der maschinellen Sprachverarbeitung nutzbringend (wieder)verwendet werden können. Damit ist auch das Thema der vorliegenden Arbeit angesprochen, die von dem Versuch handelt, exemplarisch ein Wörterbuch daraufhin zu untersuchen und zu beurteilen, inwieweit es im Rahmen der maschinellen Sprachverarbeitung von Nutzen sein könnte. Insgesamt berichte ich hier über die Ergebnisse meiner Arbeiten an der elektronischen Edition des "Oxford Advanced Learner's Dictionary", die ich in den Jahren 1990 und 1991 durchgeführt habe.

Diese Untersuchung wurde allerdings zur kritischen Stellungnahme. Es zeigte sich, daß der Aufwand, der einer Wiederverwendung der Informationen vorausgehen muß, nicht immer gerechtfertigt ist. Die meisten der verfügbaren Wörterbücher stehen in einer lexikographischen Tradition, deren Produkte Eigenschaften haben, die sie für die Wiederverwendung im Rahmen eines sprachverarbeitenden Systems ungeeignet erscheinen lassen. Die Kenntnis dieser Eigenschaften und auch die detaillierte metalexikographische Beschreibung, die als "Nebenprodukt" einer computerunterstützten Studie eines Wörterbuches entstanden ist, ist meines Erachtens für die Lexikographie und Computerlexikographie von Interesse.

Im Bereich "Computer und Lexikographie" kommen Linguisten, Lexikographen und Informatiker zusammen, die aufgrund ihrer jeweiligen ausgeprägten fachspezifischen Terminologie oft Schwierigkeiten haben, sich miteinander zu verständigen. Ich hoffe, daß ich mit diesem Buch auf der einen Seite Sprachwissenschaftlern und auf der anderen Seite Computerspezialisten gerecht werde und zwischen allen Beteiligten in verständlicher Weise vermitteln kann.

Dieses Buch verdankt sehr viel den Ratschlägen und Anregungen aus dem Kreis meiner diskussionsfreudigen Leser, denen ich an dieser Stelle ganz herzlichen Dank aussprechen möchte. Dies gilt vor allem für Ulrich Heid und Stefan Momma, die mir aus

linguistischer und informatischer Sicht mit zahlreichen Anregungen und mit viel konstruktiver Kritik zur Seite standen. Dies gilt auch für Oliver Christ, dem ich an dieser Stelle für seine scharfsinnigen Kommentare und für wertvolle Diskussionen danken möchte. Hinweise zur Sache gaben Susanne Altseimer und Bettina Bauer. Als Korrekturleser kam mir noch Michael Dorna zu Hilfe. Meinem früheren Heidelberger Dozenten Professor Dr. Herbert Ernst Wiegand bin ich für Diskussionen und für die Vermittlung der Aufnahme dieses Buches in die vorliegende Reihe sehr dankbar.

# Einleitung

In den letzten Jahren wurde immer wieder auf den Bedarf an lexikalischen Informationen in sprachverarbeitenden Systemen aufmerksam gemacht. Dabei wurde die Frage diskutiert, welche Rolle die maschinenlesbar vorliegenden Versionen (insbesondere) einsprachiger Wörterbücher als Wissensquelle für den Aufbau der Wörterbücher sprachverarbeitender Systeme spielen können. Überspitzt könnte man von dem Versuch sprechen, Wörterbücher, die für den menschlichen Benutzer geschrieben wurden, durch ein "Wörterbuch-Recycling" für die automatische Verarbeitung natürlicher Sprache wiederzuverwenden.

Insbesondere in der englischen und amerikanischen Forschungstradition finden sich Arbeiten, die sich mit einigen maschinenlesbar vorliegenden Versionen einsprachiger Wörterbücher des Englischen im Sinne einer Wiederverwendung der vorliegenden Informationen beschäftigt haben.

Im Rahmen einer für die EG-Kommision erstellten Umfrage[1] wurde festgestellt, daß immerhin die Hälfte aller lexikalischen Datenbasen in Europa, die in maschinenlesbarer From vorliegen, ausgehend von bestehenden maschinenlesbaren Versionen gedruckter Wörterbücher aufgebaut wurden. Im Kern handele es sich dabei um bekannte englische und amerikanische Wörterbücher wie das Longman Dictionary of Contemporary English (LDOCE), das Webster's Seventh New Collegiate Dictionary (W7) oder das Collins COBUILD English Language Dictionary (COBUILD).

Der Forschung für das Deutsche steht bis heute keine vollständige maschinenlesbare Ausgabe eines aktuellen und für die Wiederverwendung interessanten Wörterbuches

---

[1] Diese Umfrage wurde im Rahmen eines Forschungsauftrags der Europäischen Gemeinschaft (unter der Federführung der Universität PISA und INK International) und im Zusammenhang mit der EG-Studie EUROTRA-7 ("Feasibility and Project Definition Study on the Reusability of Lexical and Terminological Resources in Computerised Applications", Juni 1990 - August 1991) in Auftrag gegeben. Thema der EUROTRA-7 Studie war die mögliche Wiederverwendung von lexikalischen und terminologischen Ressourcen in NLP (natural language processing) und die Definition von Forschungsprojekten in diesem Bereich (vgl. [ET-7 1991]). Teilnehmer dieser Studie waren: Universität Bochum, Hachette Education, Universität Heidelberg, University of Manchester Institute of Science and Technology, Oxford University Press, Université de Paris VII, Università di Pisa, Universität Saarbücken, SEMA Group Belgium, Van Dale Lexicografie, unter Federführung der Universität Stuttgart, Institut für maschinelle Sprachverarbeitung (Koordinator der Studie: Ulrich Heid).

frei zur Verfügung. Dies ist einer der Gründe, warum dieser Arbeit eine neuere und noch relativ unbearbeitete Version eines englischen Wörterbuches zugrunde liegt. Es handelt sich um die elektronische Edition der dritten Auflage des "Oxford Advanced Learner's Dictionary of Current English"[2], die nun erstmals in größerem Umfang daraufhin untersucht wird, ob und wie (unifikationsbasierte) sprachverarbeitende Systeme von den vorliegenden Daten profitieren können.

**Das Szenarium der "Wiederverwendung"**

Die Verfügbarkeit von maschinenlesbaren Wörterbüchern (MRD, "machine readable dictionary") erfordert es, daß zwischen dem herkömmlichen gedruckten Wörterbuch, kurz dem "Papierwörterbuch", und der diesem Wörterbuch entsprechenden "elektronischen Edition" unterschieden wird. Im vorliegenden Fall nehme ich auf die gedruckte Fassung des "Oxford Advanced Learner's Dictionary" mit der Abkürzung **OALD** und auf die elektronische Edition mit der Abkürzung **OALDE** Bezug.

Seit der Einführung der modernen computergesteuerten Drucktechnik gilt für die meisten Bücher, daß ein Teil der drucktechnischen Realisierung über eine Setzmaschine abgewickelt wird, die ihrerseits durch ein "Satzband" gesteuert wird. Dieses Satzband enthält gleichzeitig den Text des Buches und die Steueranweisungen für die Setzmaschine. Auch für die meisten modernen Wörterbücher existiert ein solches Satzband, das im allgemeinen als Ausgangspunkt für die Erstellung einer elektronischen Edition herangezogen wird. Die Überführung der Daten des Satzbandes in die Daten der maschinenlesbaren Version des Wörterbuches wird mit einem sogenannten "Wörterbuchparser" bewerkstelligt, der speziell für das jeweilige Wörterbuch konstruiert wird. Der Wörterbuchparser ist ein Computerprogramm, das die Übersetzung des Satzbandes in eine maschinenlesbare Edition leisten kann. Wörterbuchverlage erstellen maschinenlesbare Editionen, um die Daten des Wörterbuches z.B. in Form eines anderen Mediums zu vertreiben[3] oder um die Daten (z.B. für den Aufbau neuer Wörterbücher) in maschinenlesbarer Form zur Verfügung zu haben. Der Übersetzungsvorgang vom Satzband in eine maschinenlesbare Edition ist aber immer der erste Schritt auf dem Weg zur Wiederverwendung der Daten, auch im Hinblick auf die Wiederverwendung im Rahmen eines sprachverarbeitenden Systems.

---

[2] Oxford Advanced Learner's Dictionary of Current English. A.S. Hornby with the assistance of A.P. Cowie, J. Windsor Lewis. 3rd ed. Oxford: Oxford Univ. Press 1974.

[3] So können Wörterbücher z.B. in Form von optischen Platten (CD-ROM) verkauft werden. Solche Wörterbücher können mit Hilfe eines entsprechenden CD-ROM-Laufwerkes benutzt werden, das an einen Computer angeschlossen ist. Immer beliebter werden auch Wörterbücher in Form eines kleinen Taschencomputers. Desweiteren werden Wörterbücher auch in Form von Softwarepaketen vermarktet, die das (meist verschlüsselte) maschinenlesbare Wörterbuch und eine spezielle Zugriffssoftware enthalten, die das schnelle Nachschlagen ermöglicht. Solche Systeme werden im folgenden Look-Up-Wörterbuch genannt.

Das Format einer maschinenlesbaren Edition eines Wörterbuches ist nicht notwendigerweise sehr gut für die weitere Bearbeitung durch Computerprogramme geeignet[4]. Wenn dies der Fall sein sollte, wird man in einem weiteren Schritt eine (möglichst) informationserhaltende Übersetzung der maschinenlesbaren Version in ein für die weitere Verarbeitung geeigneteres Format vornehmen. Diesen Vorgang nennen wir Reformatierung (vgl. [HEYN et al. 1991], [HEID et al. 1992]).

Ein sprachverarbeitendes System erwartet eine bestimmte Repräsentationsform der lexikalischen Informationen. Bei der Bemühung, diese Informationen aus einem MRD zu gewinnen, können zwei unterschiedliche Strategien verfolgt werden: Einerseits kann man sich auf einzelne Teilinformationen beschränken, die aus dem MRD herausgefiltert werden, oder man kann andererseits versuchen, größere Informationseinheiten (wie z.B. ganze Artikel) in die gewünschte Repräsentationsform des sprachverarbeitenden Systems zu überführen. Auf ersteres wird im folgenden mit dem Begriff Extraktion und auf letzteres mit dem Begriff Transformation Bezug genommen[5].

Der Transformation oder der Extraktion von Informationen muß aber eine Analyse bzw. "Reinterpretation" der lexikographischen Beschreibungen, die dem Wörterbuch zugrunde liegen, vorausgehen. Die Reinterpretation muß einerseits die Struktur der Artikel des Ausgangswörterbuches möglichst präzise erfassen; andererseits muß sie die deskriptiven Intuitionen des Autors der Ressource rekonstruieren und sollte somit detailliertes Wissen über die in den Artikeln anzutreffenden Informationen bereitstellen. Die Reinterpretation ist also eine metalexikographische Daten- und Strukturanalyse mit dem Ziel einer Beurteilung der Wiederverwendbarkeit von lexikalischen Daten aus einem MRD.

Man kann diese Schritte in ihrem Zusammenhang noch einmal wie folgt zusammenfassen: Steht eine durch einen Wörterbuchparser erzeugte maschinenlesbare Version eines Wörterbuches zur Verfügung, muß erst durch eine Reinterpretation der vorliegenden Daten festgehalten werden, wie, welche und in welchem Umfang Informationen durch Extraktionen oder Transformationen in eine für ein sprachverarbeitendes System verwendbare Repräsentationsform überführt werden können. Je umfangreicher und detaillierter diese Reinterpretation ausfällt, desto leichter wird die technische Realisierung der Extraktion oder Transformation sein. Abbildung 1 veranschaulicht nochmals diesen Zusammenhang.

---

4 Geeignete Formate wären z.B. ein Datenbankformat oder programmiersprachliche Konstrukte wie beispielsweise PROLOG-Terme oder LISP-Listen. PROLOG und LISP sind Programmiersprachen, die in der Computerlinguistik bevorzugt verwendet werden. Im vorliegenden Fall wurde das OALDE in ein für die Programmiersprache LISP verarbeitbares Format reformatiert.

5 Extraktionen bestehen aus einfachen Operationen wie "Löschen", "Ersetzen" usw., deren Ergebnis meist eine Liste "herausgefilterter" Daten ist. Transformationen setzen sich aus komplexeren, (regelgeleiteten) strukturverändernden Operationen zusammen. Beide Begriffe verstehen sich als zwei Endpunkte eines Kontinuums, zwischen denen Mischformen möglich sind.

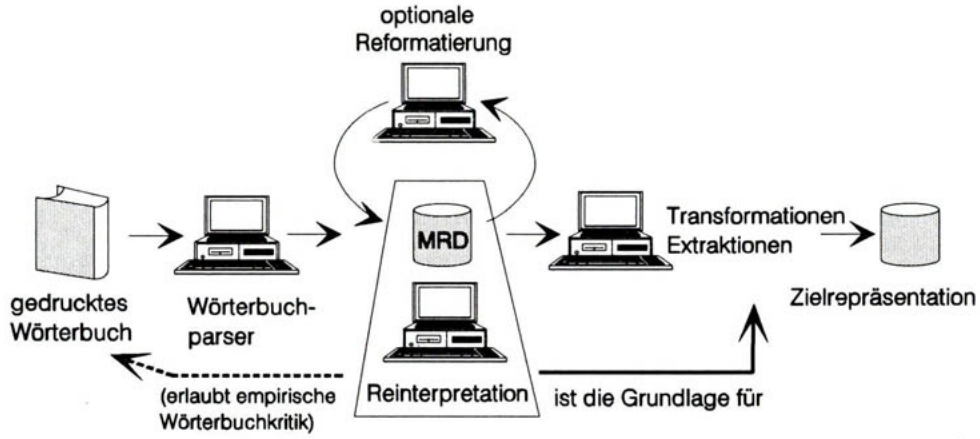


Abb. 1: Das Szenarium der "Wiederverwendung"

## Reinterpretation des OALDE

Das Ziel dieses Buches ist eine Reinterpretation des OALDE. Hieraus leiten sich die zwei grundlegenden Fragen ab, die im folgenden beantwortet werden müssen:

1. Welche strukturellen Eigenschaften der Artikel des OALDE müssen bei Extraktionen und Transformationen beachtet werden?
2. Welche linguistischen Informationen können überhaupt sinnvollerweise aus dem OALDE wiederverwendet werden?

Die Antwort auf diese Fragen wird mit Hilfe umfangreicher Analysen gegeben, die mit Computerunterstützung erstellt wurden. Außerdem werde ich mich auf metalexikographische Methoden stützen, wie sie bei der Beschreibung von Wörterbüchern in der Wörterbuchforschung entwickelt wurden.

Die Terminologie und Methoden der Metalexikographie wurden in der Computerlexikographie meines Erachtens bisher nicht genügend beachtet. Im folgenden wird der Versuch unternommen, die Methoden beider Forschungsfelder auf ein maschinenlesbares Wörterbuch anzuwenden. Dabei wird als "Seiteneffekt" auch empirische Wörterbuchkritik geübt, die "metrische" (kardinale) Aussagen über Wörterbücher bzw. über das im vorliegenden Fall zu beurteilende Oxford Advanced Learner's Dictionary möglich macht[6].

---

6 Metrische bzw. kardinale Prädikate können dann zugesprochen werden, wenn sie auf einen numerisch ausdrückbaren Sachverhalt (hier z.B. lexikographiebezogene Größen wie Fehlervorkommen, vorkommende Angabewerte usw.) zutreffen. Wiegand schreibt zum Verhältnis von metrischen Prädikaten und Wörterbuchforschung: (Vgl. [WIEGAND 1988, S. 57f]) *"Auch in der Wörterbuchfor-*

Die Qualität eines maschinenlesbaren Wörterbuches, dem ein traditionelles gedrucktes Wörterbuch zugrundeliegt, hängt natürlich entscheidend von der Beschaffenheit dieses urspünglichen Wörterbuches ab. Dieser Zusammenhang ist von großer Wichtigkeit, da die aktuellen Papierwörterbücher (noch) ein Produkt traditioneller lexikographischer Herstellungshandlungen sind. Unter dieser "traditionellen Lexikographie" verstehe ich die Wörterbuchproduktion, die durch kein - wie auch immer geartetes - computerbasiertes Lexikographiesystem zur Einhaltung einer konsistenten Form gezwungen wird. Auch andere Wörterbücher, die als Quelle von Extraktionsversuchen verwendet wurden, sind, wie das OALD, Wörterbücher der traditionellen Lexikographie. Bei diesem Wörterbuchtyp wurde von Seiten der Computerlinguisten immer wieder das Ausmaß an Inkonsistenzen bemängelt. Boguraev und Briscoe [BOG/BRI 1989, S.19] schreiben in diesem Zusammenhang:

> *These deviations mostly went undetected because of the lack of computerised checking of the structure of lexical entries (...).*

Die vorliegende Arbeit wird unter anderem diese (In-)Konsistenz für das OALDE sehr detailliert dokumentieren.

**Aufbau der Kapitel**

Kapitel 1 gibt einen Überblick über die Eigenschaften der beiden Wörterbücher OALD und OALDE. Die Wörterbücher werden hinsichtlich ihrer Stellung in einer Wörterbuchtypologie beurteilt. Dabei werden die für ein Wörterbuch charakteristischen Eigenschaften zur Sprache gebracht, wie z.B. die Auswahl der Lemmata aus der Wörterbuchbasis, die makrostrukurelle Organisation des Wörterbuches oder die Angaben die in einem Wörterbuchartikel des OALD angetroffen werden können[7]. Die vorliegende Kodierung der elektronischen Edition in einer "SGML-ähnlichen"[8] Form und in einer hier verwendeten "LISP-Notation" wird ebenfalls in diesem Kapitel beschrieben.

Kapitel 2 geht auf die hier benötigten methodischen Grundlagen aus der Computerlexikographie und der Metalexikographie ein und gibt eine Übersicht über problemati-

---

*schung sind metrische Prädikate wichtig. Die Tatsache, daß sie bisher - wenn ich es richtig einschätze - kaum angewendet wurden, spricht m.E. nicht gegen ihre Verwendbarkeit in diesem Forschungsfeld, sondern eher für eine Unterentwicklung in der metalexikographischen Begriffsbildung".*

[7] Die Wörterbuchbasis stellt die Menge aller Wörterbuchquellen dar, aus der über mehrere lexikographische Bearbeitungsprozeduren ein Wörterbuch erstellt wird. Die Auswahl der zu bearbeitenden Wörter (Lemmata) aus der Wörterbuchbasis nennt man "äußere Selektion", die Auswahl der Angaben, mit denen dann ein Lemma beschrieben wird, "innere Selektion".

[8] SGML (Standard Generalized Markup Language) ist eine standardisierte Dokumentbeschreibungssprache, die den von Systemen und Anwendungen unabhängigen Austausch von Daten ermöglichen kann [ISO 8859].

sche Eigenschaften von Wörterbüchern, die Extraktionen und Transformationen behindern.

Den zentralen Teil der vorliegenden Arbeit stellen die Kapitel 3 bis 5 dar, in denen die eigentliche Beurteilung des OALDE erfolgt.

In Kapitel 3 wird die Frage gestellt, wie der Geltungsbereich von Informationen in der Mikrostruktur des OALDE geregelt ist und welche Konsequenzen diese Regelungen für Transformationen und Extraktionen haben. Dazu wird eine Methode zur Analyse der Mikrostruktur entwickelt, die die Skopusrelationen zwischen Artikelteilen explizit beschreiben kann. Unter Skopusrelation bzw. "skopusbezogenen Relationen" versteht man in der Metalexikographie Wiegandscher Prägung die logischen Abhängigkeiten innerhalb von Textteilen eines Wörterbuchartikels [WIEGAND 1989c, S.447ff.]. In neueren Arbeiten wird dieses Problem auch unter dem Namen "Adressierungsproblematik" behandelt. Ziel der strukturellen Analyse ist eine Beurteilung, inwieweit die Integrität der Artikel bei Transformationen und Extraktionen gewahrt werden kann. Da in unifikationsbasierten sprachverarbeitenden Systemen eine Attribut-Wert-Paar-Notation linguistischer Fakten als Teil eines Repräsentationsformalismus erwartet wird, orientiert sich die hier vorgestellte mikrostrukturelle Analyse an einer Transformation der Artikel des OALDE in eine solche Attribut-Wert-Paar-Notation.

Kapitel 4 untersucht Angaben - vor allem aus den Bereichen Morphologie und Syntax - hinsichtlich ihrer Konsistenz, ihrer Verteilung und ihrer deskriptiven Adäquatheit. Standen im Kapitel 3 diejenigen Auszeichnungen des OALDE im Vordergrund, die die textuelle Struktur der Wörterbuchartikel anzeigen, sind im Kapitel 4 diejenigen Auszeichnungen von Interesse, die einem bestimmten Angabetyp zugeordnet werden können. Es wird überprüft, ob und mit welchem Ergebnis diese Angaben in Attribut-Wert-Paare überführt werden können.

Verweise sind sehr interessante Informationen eines Wörterbuches, die besonders wertvoll sind, wenn unterschiedliche Arten von Verweisen differenziert werden. Da in der elektronischen Edition zahlreiche Auszeichnungen für die Verweise vorgefunden werden können, geht das Kapitel 5 der Frage nach, ob diese reichhaltig klassifizierten Verweise z.B. an paradigmatische Relationen wie Synonymie oder Antonymie gebunden werden können.

**Betrachtungsebenen**

Es wird im Verlauf des Buches immer wieder auf drei Ebenen argumentiert werden: einmal auf der Ebene der elektronischen Edition (OALDE), die bei allen Betrachtungen im Vordergrund steht, zweitens auf der Ebene des Papierwörterbuches (OALD) und drittens auf der Ebene des "Wörterbuchparsers", der die Artikel des OALD in die Artikel des OALDE übersetzt hat. Ich habe versucht - soweit dies möglich und not-

wendig war - diese drei Ebenen zu trennen und gegebenenfalls die Abhängigkeiten zu beschreiben.

**Die vierte Auflage**

Leider ist die vierte Auflage des Oxford Advanced Learner's Dictionary von 1989 (OALD4) nicht in einer maschinenlesbaren Edition verfügbar. Eine entsprechende elektronische Edition wurde allerdings von Oxford University Press angekündigt. Da die Unterschiede zwischen der dritten und vierten Auflage des Oxford Advanced Learner's Dictionary (OALD4) gravierend sind, werden wichtige Änderungen - soweit sie anhand des Papierwörterbuches der vierten Auflage überblickt werden können - immer wieder zur Sprache kommen.

**Anhänge**

Da viele Artikel auch in der Fassung der elektronischen Edition zitiert werden, sind in den Anhängen die wichtigsten Informationen zur elektronischen Edition in Form von Tabellen wiedergegeben. Anhang A gibt eine Tabelle über Namen und Häufigkeit der Auszeichnungen wieder. Anhang B hält die Kodierungen der Transkriptionszeichen fest, die bei der Angabe zur Aussprache Verwendung finden. Anhang C gibt die Kodierung spezieller Sonderzeichen wieder.

Um den Umgang mit diesem Buch zu vereinfachen, wurde noch ein Verzeichnis hinzugenommen (Anhang D), in dem für jede Auszeichnung der elektronischen Edition festgehalten ist, in welchem Kapitel sie behandelt wird. Anhang E besteht aus einem Verzeichnis der Tabellen und Abbildungen. Ferner wurden alle Lemmata der zitierten Artikel in den Index aufgenommen.

# Kapitel 1

# Das Papierwörterbuch und die elektronische Edition

Gedruckte Wörterbücher und die lexikographischen Handlungen, die ihrer Herstellung zugrundeliegen, sind Forschungsgegenstand der Metalexikographie. Bei dem Versuch, über Wörterbücher zu reden, wurden zahlreiche Termini geprägt, die zum Teil jedoch verwirrend und manchmal auch widersprüchlich von den unterschiedlichen Autoren verwendet werden. Bei der folgenden Beschreibung der Eigenschaften der gedruckten Fassung des OALD wird daher fast ausschließlich die elaborierte Terminologie aufgenommen, die von Hausmann und Wiegand in [HAU/WIE 1989] eingeführt wurde. Bei der Arbeit von Hausmann und Wiegand handelt es sich um einen Referenzaufsatz, der einen wesentlichen Teil der benötigten metalexikographischen Terminologie in den Sprachen Englisch, Deutsch und Französisch vorschlägt.

Bis zu einem gewissen Grad kann mit dieser Terminologie auch über die elektronische Edition geredet werden, wobei einige Modifikationen notwendig werden, die im folgenden auch entsprechend gekennzeichnet werden.

## 1.1 Die Aufbereitung des OALDE

Das zu der elektronischen Edition mitgelieferte Handbuch[1] [OALDE 1989] bezieht sich auf eine maschinenlesbare Version ('Expanded Version') von 1974 - also auf das Satzband des gedruckten Wörterbuches - aus der die in dieser Arbeit benutzte SGML-ähnliche Version aufbereitet wurde[2].

---

[1] Das Handbuch beschreibt die Auszeichnungen der elektronischen Edition, macht einen kurzen Vergleich zwischen diesen und denen des gedruckten Wörterbuches, listet die speziellen Kodierungen für Sonderzeichen (z.B. Transkription der Aussprache) auf und dokumentiert die Such-Software PAT, die von Oxford University Press mitgeliefert wird. Alle Informationen, die für die Benutzung der elektronischen Edition des OALDE notwendig sind, werden aber auch in der vorliegenden Arbeit in Form von Tabellen im Anhang wiedergegeben.

[2] Einzelheiten zur SGML-Kodierung sind im Kapitel 1.5 zu finden.

Ein Vergleich des 26. Drucks (1987) mit der 1. Druckausgabe der dritten Auflage (1974) zeigt geringfügige Unterschiede, teils in den Artikeln, teils durch weggefallene oder neue Lemmata (Stichwörter). Man kann davon ausgehen, daß die Änderungen, die seit der 1. Druckausgabe 1974 in den folgenden Auflagen vorgenommen wurden, nicht in der elektronischen Fassung zu finden sind, wie einige von mir durchgeführte Stichproben zeigen.

Bei der maschinenlesbaren Ausgangsform von 1974 handelt es sich um das Satzband des Papierwörterbuches, also der Datei, die ursprünglich für die Steuerung der Setzmaschine und somit für die drucktechnische Realisierung bestimmt war. Die Aufbereitung dieses Satzbandes in die vorliegende Form des OALDE erfolgte mit einem "Wörterbuchparser" auf der Basis dieses Satzbandes.

Unter einem Wörterbuchparser versteht man im Bereich der Computerlexikographie ein Computerprogramm, dessen Eingabe ein Satzband mit allen gerätespezifischen Druckanweisungen ist und dessen Ausgabe eine Datei ist, in der diese Druckanweisungen durch sinnvolle (standardisierte) Auszeichnungen ersetzt werden. Die Entwicklung eines solchen Parsers ist meist aufwendig, da sich die vorzunehmenden Ersetzungen im allgemeinen recht kompliziert gestalten. Oft treten Situationen ein, in denen maschinell gar nicht oder nur sehr umständlich zu entscheiden ist, wie die Daten des Satzbands interpretiert werden müssen.

Mit Hilfe von zwei Beispielen soll dieser Übersetzungsvorgang des Wörterbuchparsers einmal näher betrachtet werden.

Angenommen, in einem Wörterbuch erscheint eine Angabe zur Wortart immer unmittelbar auf das Lemma folgend und in kursiver Schriftlage (*Italique*). In diesem Fall hätte ein Wörterbuchparser gleich zwei Anhaltspunkte zum Erkennen der Angabe zur Wortart: einen positionellen und einen typographischen. Dabei ist zu beachten, daß die typographischen Auszeichnungen des gedruckten Wörterbuches im Satzband als nicht-typographische Steueranweisungen für die spezifische Druckmaschine erscheinen. Ein Wörterbuchparser würde folglich in eine mnemotechnisch sinnvolle Auszeichnung, wie z.B. "Wortart:" oder "Kategorie:" oder "part_of_speech:" oder "pos:" übersetzen, der dann der eigentliche Angabetext (z.B. "Verb") folgen würde. Eine so entstandene nicht-typographische Auszeichnung der elektronischen Edition wird - der englischen Terminologie folgend - "Tag" genannt. Der Angabetext, im folgenden meist "Angabewert", ist immer von einem bestimmten Typ, wie hier vom Typ "Angabe zur Wortart". Angaben haben also immer einen bestimmten Typ und einen bestimmten Wert[3]. Der Angabetyp sollte idealerweise im gedruckten Wörterbuch durch positionelle oder typographische (oder auch symbolische) Regelungen eindeutig erkennbar sein. In einer elektronischen Edition werden die Angabetypen mittels der

[3] Zum Terminus Angabe vgl. [WIEGAND 1989c, S. 427ff.]

Tags gekennzeichnet. Die korrekte Übersetzung der einen Form in die andere ist Sache des Wörterbuchparsers.

Ein weiteres Beispiel bietet die Gelegenheit, auf einige wesentliche Eigenschaften von Wörterbüchern zu sprechen zu kommen.

Im OALD finden sich zwei Diakritika, die große Box □ und die kleine Box ▫ , deren Funktion darin besteht, den Benutzer auf bestimmte Textteile im Wörterbuchartikel zu leiten. Man spricht hier in der metalexikographischen Terminologie von "Leitelementen" eines schnellen Zugriffspfades innerhalb eines Wörterbuchartikels. Solche Diakritika zeigen dem Benutzer - in nicht-typographischer Weise - die Struktur des Artikels an (man nennt sie deshalb auch "nicht-typographische Strukturanzeiger").

Die große Box leitet einen Textteil des Artikels ein, der das "Hauptlemma" (das eigentliche Stichwort zu Beginn eines Artikels) unter einer anderen Wortart behandelt, als dies zu Beginn des Artikels bzw. vor dem neuen Textabschnitt der Fall war. Die kleine Box leistet das gleiche für sogenannte "Subartikel". Ein Subartikel ist ein Textteil des Artikels, der ausschließlich einem Untereintrag, dem "Sublemma" zugeordnet ist. Man spricht auch davon, daß dieser Textteil an das Sublemma "adressiert" ist. Beispiele für Sublemmata sind z.B. Komposita (Zusammensetzungen) oder Derivata (Ableitungen). Dem Benutzer wird also mit der kleinen und der großen Box ein Wortartwechsel signalisiert[4]. Abb. 2 gibt einen schematischen Überblick über diesen Sachverhalt.

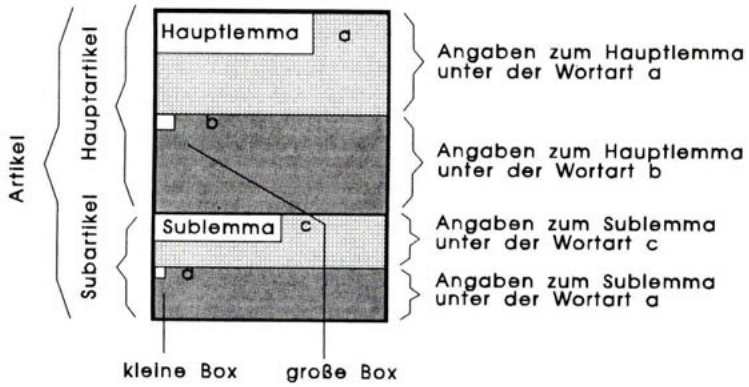


Abb. 2: Die verschiedenen Textteile eines Wörterbuchartikels: Hauptartikel und Subartikel. Die große und die kleine Box trennen im OALD Textteile, deren Angaben sich auf das jeweilige Lemma in einer bestimmten Wortart a, b oder c beziehen

[4] Die "Box" trennt also Textteile des Wörterbuchartikels, weshalb sie auch als "Separator" bezeichnet wird.

Ein Beispiel findet sich im folgenden Artikel zum Lemmazeichen *sand*, der auf den hier interessierenden Sachverhalt gekürzt wurde[5]:

> **sand** /sænd/ *n* **1** [U] (mass of) fine crushed rock as seen on the seashore, in river-beds, deserts, etc:... □ *vt* [VP6A] cover, sprinkle or scrub with ~ . **-y** *adj* (-ier, -iest) **1** covered with or consisting of ~ ... □ *n* (colloq) nickname given to a person with yellowish-red hair.

Dem Hauptlemma *sand* sind zwei Textteile zugeordnet, die durch die große Box getrennt sind. Im ersten hauptlemmatisch adressierten Textteil sind Angaben zu finden, die *sand* als Nomen beschreiben, und im zweiten Textteil wird *sand* als transitives Verb beschrieben. Das gleiche gilt für den eingebetteten Subartikel zum Derivat *sandy*, in dem einmal das Sublemma als Adjektiv und einmal als Nomen beschrieben wird.

In diesem Zusammenhang ist auch auf die Funktion des Hauptlemmas hinzuweisen, den Benutzer über die Alphabetisierung auf den Artikel hinzuleiten. Daher fungiert das Hauptlemma im allgemeinen als "Leitelementträger" der Zugriffsstruktur eines gedruckten Wörterbuches. Diese "äußere Zugriffsstruktur" über die alphabetisch angeordneten Hauptlemmata wird traditionell mit "Makrostruktur" bezeichnet, während die textuelle Strukturierung der Informationen innerhalb des einzelnen Wörterbuchartikels in der klassischen Terminologie "Mikrostruktur" genannt wird[6]. Eine vereinfachte Darstellung dieses Sachverhalts gibt die folgende Abbildung wieder:

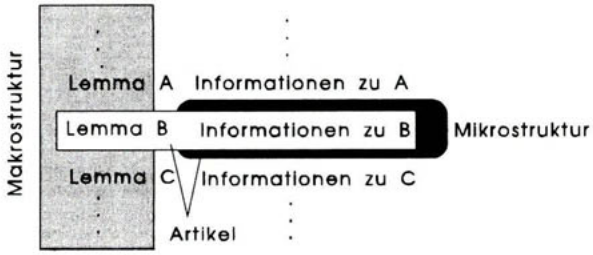


Abb. 3: Eine vereinfachte grafische Darstellung der Termini Makro- und Mikrostruktur

Die mikrostrukturelle Gliederung des Ausschnitts aus dem Artikel zum Lemmazeichen *sand* ist bereits relativ komplex. Der eingebettete Subartikel ist - strukturell gesehen - ein identisches (rekursives) Abbild des Hauptartikels. Die Angaben innerhalb der Mi-

---

5 Artikel des Papierwörterbuches werden in einer Form zitiert, die der Typographie des originalen Artikels nachempfunden ist. Für die in dieser Arbeit interessierenden Aspekte reicht diese Zitierform aus.

6 Zu den Begriffen Makro- und Mikrostruktur und deren spezifischen Probleme vgl: [WIEGAND 1989 a,b,c].

krostruktur haben nicht nur einen Typ und einen Wert, sondern haben auch einen bestimmten Bezugspunkt im Artikelganzen. Man sieht außerdem, daß die vorliegende Mikrostruktur eine Hierarchie von Textteilen bildet.

Für die beiden Diakritika "kleine Box" und "große Box" sind im OALDE explizite Auszeichnungen bzw. Tags eingeführt worden: `hps` [head part-of-speech section] und `cps` [compound part-of-speech section][7]. Diese Art der Auszeichnung mittels Tags nennt man auch "Markup", wobei sich dieser Terminus nur auf die Auszeichnung von Textteilen in maschinenlesbaren Texten bezieht. Für die Instanzen dieser Auszeichnungen wie die Tags `hps` und `cps` findet sich manchmal auch die Bezeichnung "Label" (mehr zu Markups, Tags und Labels im Abschnitt 1.5.).

Es ist also naheliegend, daß ein Wörterbuchparser bei jedem Vorkommen der Drucksteuerzeichen für die jeweiligen Diakritika der kleinen und großen Box systematisch in die entsprechenden Tags `hps` und `cps` übersetzt.

Wichtig ist in diesem Zusammenhang aber auch, daß diese Übersetzung im vorliegenden Fall zwar automatisch, jedoch nicht **eineindeutig** erfolgt ist, da sich zahlreiche Fälle finden lassen, in denen die durch Markup realisierten Strukturen der elektronischen Fassung nicht das wiedergeben, was in der gedruckten Fassung intendiert war. Dabei handelt es sich meistens um fehlerhafte typographische Auszeichnungen des gedruckten Wörterbuches oder um Fehler des Wörterbuchparsers. In den Kapiteln 3 bis 5 werden zahlreiche Beispiele für das Auftreten solcher Fehler diskutiert.

## 1.2 Wörterbuchtyp und Typ des maschinenlesbaren Wörterbuches

In Anlehnung an die Wörterbuchtypologie von [HAUSMANN 1989a] kann man das OALD zu den allgemeinen einsprachigen Wörterbüchern bzw. den Definitionswörterbüchern in der Ausprägung der sog. "learner's dictionaries" zählen [HAUSMANN 1989b, S. 982]. Es steht damit in einer Reihe mit Wörterbüchern wie etwa dem COBUILD oder dem LDOCE. Das OALD ist ein einsprachiges Wörterbuch der englischen Standardsprache mit Schwerpunkten im Bereich der Bedeutungsangaben und vielfältigen Angaben zur englischen Grammatik. Eine Anzahl von lexikalischen und grammatikalischen Schwierigkeiten des Englischen wird für den Adressatenkreis der Englisch lernenden Nicht-Muttersprachler aufbereitet. Dieser didaktische Anspruch ist das Leitmotiv des federführenden Herausgebers A.S. Hornby, der im Schutzumschlag der gedruckten Ausgabe die folgenden Argumente für sein Werk anführt:

---

[7] Eine Liste aller Tags findet sich im Anhang A.

- einfache Begriffsdefinitionen (*"practical definitions and usage notes in simple English"*) und Beispielsätze;
- Aufbereitung der *idioms*[8] und *phrasal verbs*;
- zahlreiche Informationen zur Morphologie und Syntax: Irreguläre Verbformen, Plurale und Komparation, *verb patterns*, *countable* und *uncountable* Markierungen;
- Transkriptionen nach dem internationalen phonetischen Alphabet (IPA);
- Illustrationen, stilistische Hinweise, Unterschiede zwischen britischem und amerikanischem Englisch und zahlreiche Anhänge im Wörterbuchnachspann.

Die Einordnung des OALDE als MRD ist schwieriger, da sich eine Typologie der maschinenlesbaren lexikalischen Ressourcen erst formiert. Das liegt zum einen daran, daß die Computerlexikographie ein junges, sich rasch entwickelndes Fachgebiet ist und in dieser Entwicklungsphase terminologische Unstimmigkeiten noch relativ schwer zu überblicken sind[9]. Zum anderen ist die Gefahr noch groß, daß Fallbeispiele vorschnell zu prototypischen Vertretern einer Klasse avancieren, obwohl sie nur durch heterogene Klassifikationskriterien beschreibbar sind. In dieser Situation ist es sinnvoller, wenn die relevanten Klassifikationskriterien erst einmal sorgfältig zusammengestellt werden[10]. Wichtige Vorschläge zur Klassifikation betreffen die Repräsentationsform und den Formalitätsgrad (auch Formalisierungspotential) der maschinenlesbaren Ressourcen [MAR/WOL 1989]. Ebenso müssen die den Ressourcen zugrundeliegenden Quellen bei einer Klassifikation einbezogen werden[11].

Eine Sammlung solcher Kriterien wurde im Rahmen der EUROTRA-7 Studie zusammengestellt (vgl. [HEID et. al. 1990]). Mit Hilfe dieser Kriterien läßt sich das OALDE besser beschreiben als durch eine Einordnung in eine der bestehenden tentativen Klassifikationen. Sechs Hauptkriterien werden aufgeführt: Anwendungsperspektive der Ressource, inhaltliche Parameter, organisatorische Parameter, technische Parameter, der Aspekt der Abhängigkeit von zugrundeliegenden Quellen sowie kommerzielle und juristische Parameter. Auf das OALDE angewandt, kann man dann folgendes formulieren:

---

8 Termini der englischen Grammatik, die nicht übersetzt wurden, sind als solche im Text typographisch hervorgehoben.

9 So existieren nebeneinander Begriffe wie: Machine Readable Lexicon, Machine Readable Dictionary (MRD), Machine Dictionaries, Lexical Data Bases (LDB), Large-scale lexical database (LDB), Computer-based dictionaries, lexical term banks usw.

10 Ein solches Vorgehen führt zu Matrizen, die als Ausgangspunkt für explizite Typologisierungen dienen können (vgl. [WIEGAND 1988, S. 91]).

11 Vgl. die tentativen Klassifikationen von [CALZOLARI 1989, S.510-512] und [CALZOLARI et al.1987, S. 66-67]).

### Anwendungsperspektive der Ressource

Ein besonderer Zweck wurde mit dem OALDE außerhalb der verlagsinternen Verwendung bei Oxford University Press nicht verbunden. Vorerst wird die elektronische Edition im Zusammenhang mit dem Problem "Wiederverwendung lexikalischer Ressourcen in sprachverarbeitenden Systemen" durch die Forschung bearbeitet. Eine mögliche Anwendung ist natürlich, mittels geeigneter Programme, die Benutzung als elektronisches Wörterbuch, das über einen Bildschirm benutzt werden kann. Dies könnte in Form eines CD-ROM-Wörterbuches, als Look-Up-Wörterbuch oder in Form eines kleinen Taschencomputers erfolgen[12].

### Inhaltliche Parameter

Die inhaltlichen Parameter des OALDE fallen mit denen des OALD zusammen, da die elektronische Edition in dieser Hinsicht dem Papierwörterbuch weitgehend entspricht. Eine genauere Untersuchung dieser Informationen ist Teil der vorliegenden Arbeit. Für die Belange einer Typologie stehen die obigen Ausführungen zur Wörterbuchtypologie.

### Organisatorische Parameter

Es existiert explizite Information neben impliziter Information, entsprechend den lexikographischen Informationen des Papierwörterbuches. Die textuellen Verdichtungstechniken wie die Tilde (~) oder Teilstringangaben (man spricht auch von lexikographischer Textkondensierung), finden sich ebenso im OALDE wieder.

In organisatorischer Hinsicht ist die "SGML-ähnliche" Kodierung am bemerkenswertesten[13]. SGML-kodierte Dokumente bestehen aus dem mit SGML-Markup versehenen Dokument und einer sogenannten Dokumenttypdefinition (DTD), die in Form einer Grammatik das Aussehen des Dokumentes beschreibt. Das bedeutet unter anderem, daß festgehalten wird, welche Tags erlaubt werden, welche Abfolgen und Einbettungen von Tags möglich sind u.a.m. Diese Dokumenttypdefinition fehlt bei dem OALDE, weshalb hier auch nur von einer SGML-ähnlichen Kodierung gesprochen wird.

---

[12] Der Nutzen einer solchen Anwendung sollte nicht unterschätzt werden. Die Look-Up-Wörterbücher von Gyldendal, z.B. das "Gyldendals elektronisk ordbøg Dansk-tysk" (GEO-DT) können hier als relativ früh verfügbare kommerzielle Produkte genannt werden [vgl. auch HEYN 1990a]. Oxford University Press bietet beispielsweise das Oxford English Dictionary (OED2) auf Band oder auf CD-ROM an. Unter anderem wurde auch von mir ein speicherresidentes Programm unter DOS geschrieben, das das OALDE als Look-up-Wörterbuch zur Verfügung stellt.

[13] Abschnitt 1.5 behandelt die SGML-Kodierung des OALDE im Detail.