Non-thematic Part

Carolin Müller-Spitzer

Ordnende Betrachtungen zu elektronischen Wörterbüchern und lexikographischen Prozessen

„Es gibt nichts rundum Zutreffendes, Eindeutiges und Stichhaltiges, das ich über mich sagen, gar ohne Wenn und Aber in einem einzigen Wort ausdrücken könnte. Ich *unterscheide*, dies ist das A und O meiner Logik."
Michel de Montaigne, Essais

1 Zum Thema dieses Aufsatzes

Die Wörterbuchlandschaft hat sich durch das Hinzukommen des elektronischen Mediums in den letzten ein bis zwei Jahrzehnten massiv verändert.[1] Waren es erst zaghafte Versuche, gedruckte Wörterbücher in sehr ähnlicher Form einfach elektronisch verfügbar zu machen, so steht dem heute eine sehr große Vielfalt an elektronischen Wörterbüchern, Wortschatzinformationssystemen – oder wie immer man sie auch bezeichnen mag – gegenüber. Das elektronische Medium spielt darüber hinaus an den verschiedensten Stellen im lexikographischen Prozess eine Rolle, ob bei der computerunterstützten Erstellung gedruckter Wörterbücher, bei der Datengewinnung aus elektronischen Textkorpora oder als Publikationsmedium bei elektronischen Wörterbüchern. Auch wenn die verfügbaren elektronischen Wörterbücher hinter den eigentlichen Möglichkeiten des elektronischen Mediums für die Lexi-

1 Ich danke INGRID SCHMIDT für die wie immer wertvollen und anregenden Diskussionen zu diesem Thema.

kographie zurückbleiben[2] – wie es die Wörterbuchforschung immer wieder feststellt – stellen diese neuen Produkte schon jetzt eine Herausforderung an die Wörterbuchforschung dar. Vor allem die Wörterbuchkritik hat hier ein stark erweitertes Feld vor sich und Fragen zu beantworten wie: Können Bewertungsmaßstäbe, die sich für gedruckte Wörterbücher herausgebildet haben, auf elektronische Wörterbücher übertragen werden? Und wenn dies gelten soll, gilt das für alle Arten elektronischer Wörterbücher? Oder wie soll man elektronische Nachschlagewerke klassifizieren, damit hier besser differenziert werden kann? Ist die automatische Gewinnung von Daten aus elektronischen Textkorpora als lexikographische Tätigkeit zu bezeichnen? Was ist mit der Verwendung lexikographischer Daten in der maschinellen Sprachverarbeitung? Gehört letzteres auch zu Themen der Wörterbuchforschung?

In diesem Aufsatz sollen einige dieser Fragen aufgegriffen werden. Dabei wird v.a. die Perspektive der Wörterbuchkritik, d.h. die Frage nach der Bewertung und Einordnung elektronischer Wörterbücher eingenommen. Die hier dargestellten Überlegungen gliedern sich in drei Themenbereiche mit unterschiedlicher Gewichtung: Kapitel 2 beschäftigt sich Unterscheidung von automatisch erstellten vs. menschlich bearbeiteten elektronischen Wörterbüchern und den Herausforderungen, die so entstehende Produkte an die Wörterbuchkritik stellen. Warum gerade diese Art der Wörterbucherstellung eine besondere Schwierigkeit für die metalexikographische Kritik darstellt, wird an einem vorangestellten Ausgangsbeispiel veranschaulicht. In diesem Zusammenhang wird auch der Versuch von terminologischen Klärungen in Bezug auf elektronische Wörterbücher unternommen. Der zweite Themenbereich zeigt in einer Gesamtschau Arten lexikographischer Prozesse, wie sie heute in der Wörterbuchlandschaft zu finden sind. Auch hier wird ein Beispiel vorangestellt, welches veranschaulichen soll, wie heute lexikographische Prozesse aussehen können und in welche Richtung demnach ein Gesamtüberblick über Arten von lexikographischen Prozessen erweitert werden muss. Kapitel 4 greift ein Thema heraus, welches besonders für den medienneutral konzipierten lexikographischen Prozess (zu diesem Terminus s. 3.3) und den computerlexikographischen Prozess relevant ist: eine mögliche Sicht der Ebenen im lexikographischen Prozess. Alle drei Themenbereiche bieten Einordnungsmöglichkeiten für die Wörterbuchkritik. Dabei sind sie nicht für alle Arten elektronischer Wörterbücher gleichermaßen relevant.

Um es deutlich herauszustellen: Es geht in diesem Aufsatz insgesamt nicht um einzelne Kriterien, nach denen elektronische Wörterbücher bewertet werden können. Vorschläge dazu sind u.a. in LEHR (1996), KLOSA (2001), und ENGELBERG/LEMNITZER (2001), bes. 194ff. genannt. Die hier entwickelten terminologischen Unterscheidungen verstehen sich als Ergänzung zu diesen Arbeiten in dem Sinne, dass ein elektronisches Wörterbuch – bevor es im einzelnen bewertet wird – nach den hier zu entwickelnden Unterscheidungen eingeordnet werden kann. Insgesamt handelt es sich demnach um Betrachtungen, die helfen sollen, den neuen und sehr vielfältigen Gegenstandsbereich der elektronischen Wörterbücher und lexikographischen Prozesse aus verschiedenen Perspektiven zu ordnen.

2 Für diese Meinung wären zahlreiche Beispiele zu nennen. Verwiesen sei an dieser Stelle lediglich auf die Rezensionen der Lexicographica in der Rubrik „Electronic dictionaries“ und exemplarisch auf ENGELBERG/LEMNITZER (2001), bes. 222ff., STORRER (2001) und FELDWEG (1997).

2 Arten elektronischer Wörterbücher

2.1 *Müller*, *Edelmüller* und *Mehlkäfer*

Die automatische Analyse elektronischer Textkorpora macht die Entwicklung ganz neuer Produkte möglich. Dies ist ein neuer Gegenstandsbereich für die Wörterbuchforschung, da die Möglichkeit, überhaupt solche wortschatzbezogenen Daten in der Weise *automatisch* erstellen zu können, durch die umfassende Anwendung des neuen Mediums bedingt ist. Hinzu kommt, dass durch das Internet eine sehr preiswerte Publikationsmöglichkeit zur Verfügung steht, die zudem eine sehr hohe Verbreitung garantiert. Dadurch kommen Produkte auf den Markt, die in der Weise oder in der Form nicht gedruckt worden wären. Welche neuen Fragen dies für die Wörterbuchkritik aufwerfen kann, soll am folgenden Beispiel, dem Projekt DEUTSCHER WORTSCHATZ der Universität Leipzig (WORTSCHATZ UNI LEIPZIG), demonstriert werden.

Es handelt sich nach Projektinformationen hierbei um ein „umfangreiches Vollformenwörterbuch des Deutschen“ (QUASTHOFF/WOLFF 1999, 1), in dem „die typischen Inhalte und Funktionen unterschiedlicher Wörterbuch- und Lexikontypen […] zur Verfügung stehen (Nachschlagen von Begriffen; Querverweise; morphologische, syntaktische, semantische und pragmatische Information, statistische Daten; Einarbeitung von Ontologien) und durch die zusätzlichen Möglichkeiten des elektronischen Mediums (automatische linguistische Analyseverfahren, Recherche, Hypertextualisierung, automatische Generierung unterschiedlich strukturierter Einträge, Visualisierung von Relationen zwischen Einträgen) ergänzt werden“ (QUASTHOFF/WOLFF 1999, 1). Auf der Webseite wird das Produkt auch allgemein als ein „Nachschlagewerk für Wörter und ihren Gebrauch bezeichnet“ (WORTSCHATZ UNI LEIPZIG).

Interessiert man sich nun beispielsweise für seinen Nachnamen, so wie ich hier für „Müller“, so erhält man den in Abbildung 1 gezeigten Eintrag.

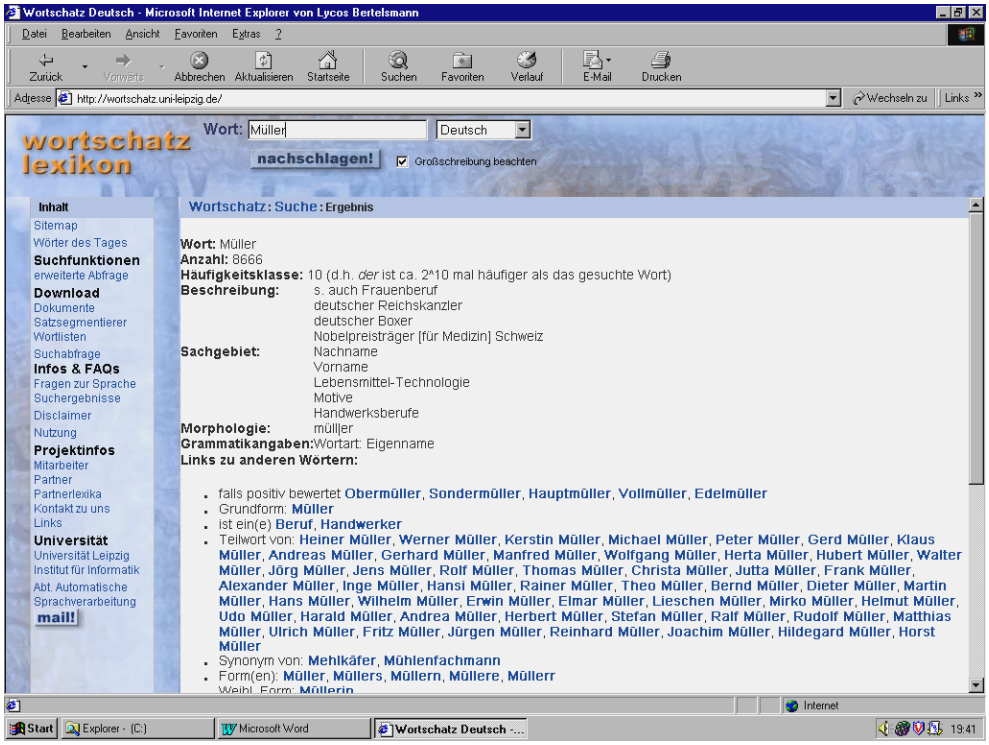

Abb. 1: Eintrag „Müller“ aus dem WORTSCHATZ-LEXIKON (gesehen am 1. Juni 2003)

Die Lemmatisierung von Eigennamen kann für allgemeine, einsprachige Wörterbücher als unüblich angesehen werden, so heißt es z.B. in der Einleitung zum DUDEN-UNIVERSALWÖRTERBUCH: „Personennamen“ wurden nur „ aufgenommen, wenn sie als Appellativa (Gattungsbezeichnungen) […] gebraucht werden, z.B. Casanova“ (DUDEN-DUW3, 7). Diese Regel kann als gängige Praxis angesehen werden. Im DUDEN-UNIVERSALWÖRTERBUCH findet sich dementsprechend auch unter „Müller“ nur die Erläuterung der Berufsbezeichnung. Das WORTSCHATZ-LEXIKON wird von den eigenen Mitarbeitern, wie oben zitiert, jedoch als „Vollformenwörterbuch“ bezeichnet; wobei anscheinend Eigennamen zu den Vollformen gerechnet werden.

Wirklich überraschend wird es aber unter der Rubrik „Links zu anderen Wörtern“. Zum Beispiel steht da zu „Müller“: „falls positiv bewertet: Obermüller, Sondermüller, […] Edelmüller“. Verfolgt man diese Angabe und klickt z.B. auf „Edelmüller“, so erhält man den in Abbildung 2 gezeigten Eintrag.

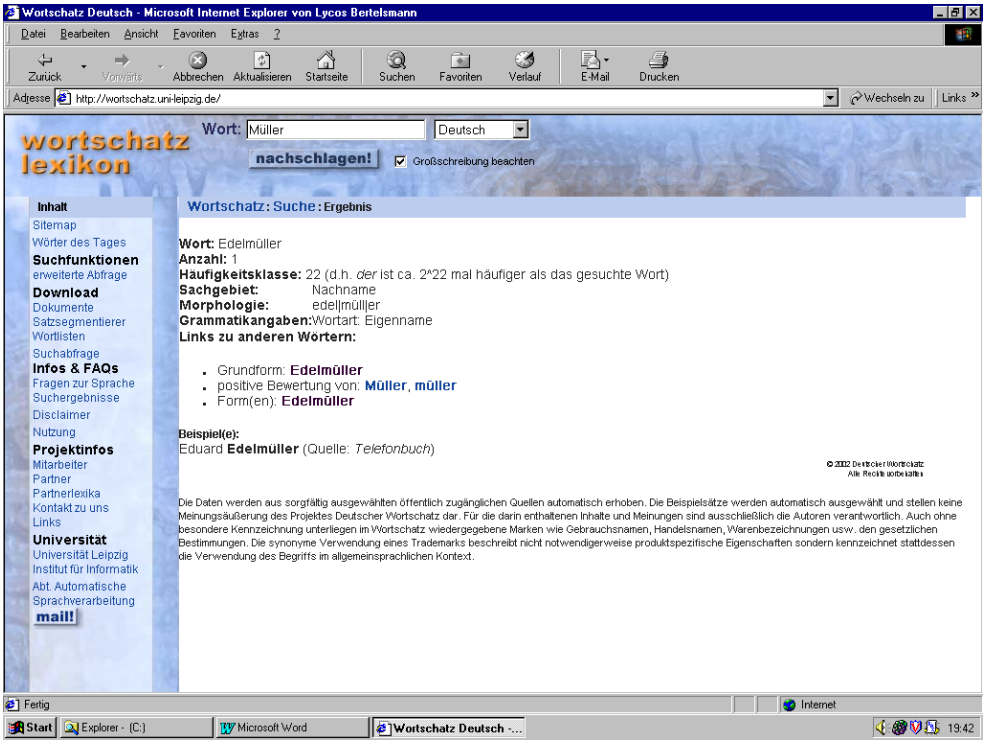

Abb. 2: Eintrag „Edelmüller“ aus dem WORTSCHATZ-LEXIKON (gesehen am 1. Juni 2003)

Als Sachgebiet für „Edelmüller“ wird „Nachname“ angegeben, das einzige Beispiel, was aufgeführt wird, stammt aus dem Telefonbuch, welches anscheinend in das Korpus aufgenommen wurde.

Doch zurück zu „Müller“, denn dort erwartet uns noch mehr Beachtenswertes: Müller ist nicht nur synonym zu „Mühlenfachmann“, sondern auch zu „Mehlkäfer“. Wie es zu dieser Vernetzung kommt, ist nicht zu erkennen.[3]

Dieses einzeln herausgegriffene, bewusst strittige Beispiel soll in diesem Zusammenhang zeigen: Wollte man das WORTSCHATZ-LEXIKON rezensieren und würde ganz allgemein die Bewertungsmaßstäbe, die sich für gedruckte Wörterbücher etabliert haben, übertragen[4], dann müsste diese Kritik verheerend ausfallen. Denn die Vernetzung, die oben gezeigt wurde, ist unter lexikographischen Gesichtspunkten nicht nachvollziehbar. Aber ist es überhaupt legitim, diese Anforderungen an das WORTSCHATZ-LEXIKON zu stellen? Dafür ist es interessant, den Passus, der unter jedem Eintrag steht, einmal näher zu betrachten.

3 Angesprochen auf diese oben gezeigte Vernetzung kündigte STEFAN BORDAG, ein Mitarbeiter des Leipziger Projektes, auf einem Kolloquium in Mannheim am 13. März 2003 (zu dem Kolloquium s. `www.ids-mannheim.de/aktuell/jt2003/kolloquium2003.html`>) an, dass dieses „Problem“ in einer neuen Version des WORTSCHATZ-LEXIKONS behoben sein werde.

4 In diesem Sinne z.B. ENGELBERG/LEMNITZER (2001), 194: „Auch für elektronische Wörterbücher gelten deshalb zunächst die Bewertungskriterien und -maßstäbe, die wir weiter oben für Printwörterbücher aufgezählt und kommentiert haben.“

Dort heißt es: „Die Daten werden aus sorgfältig ausgewählten öffentlich zugänglichen Quellen automatisch erhoben […].“ Es gilt also bei den Einträgen nicht – wie sonst üblich – zu fragen: Wie kommen die Lexikographen dazu, diese Angabe in der Form anzusetzen? Sondern vielmehr müssen die Fragen in folgende Richtung gehen: Mit welchen Analysemethoden werden die Textkorpora analysiert? Wie werden die Angaben im Detail automatisch erstellt? Wie ist das zugrunde gelegte Textkorpus zusammengesetzt? Ist diese Zusammensetzung nachvollziehbar? Welche Ziele verfolgen die Projektverantwortlichen mit ihrem Produkt? etc. Die Fragen stimmen also zum großen Teil vom Ansatzpunkt nicht mit denen überein, die man üblicherweise bei einer Wörterbuchrezension stellt (vgl. z.B. ENGELBERG/LEMNITZER 2001, 161ff.).

Leider finden sich zu der Gewinnung der einzelnen Angaben auf der Webseite relativ wenig Hinweise, jedoch schreiben WOLFF und QUASTHOFF: „Die Angaben zu Flexion, Morphologie, Sachgebiet, Synonymie etc., die vor allem aus strukturierten Quellen wie Lexika und durch computerlinguistische Analyseprogramme ermittelt werden, sind der jeweiligen Grundform eines Wortes zugeordnet […]. Momentan haben diese Angaben noch nicht überall die gewünschte Qualität, da sie teilweise mit automatischen Mitteln erzeugt wurden und bei der Anzahl der Einträge eine vollständige intellektuelle Überarbeitung nicht möglich ist.“ (QUASTHOFF/WOLFF 1999, 2) Aus diesem Grund sollte man natürlich auch vorsichtig mit der Behauptung sein, das Wortschatz-Lexikon erfülle die Funktionen eines üblichen Wörterbuchs, denn zu denen gehört – zumindest im Rahmen der wissenschaftlichen Lexikographie (vgl. WIEGAND 1998, 40ff.) – auch die Verantwortung für die Zuverlässigkeit der lexikographischen Daten (vgl. WIEGAND 1997, 195). Doch allein vom Produkt her kann man den Anspruch an eine solche Zuverlässigkeit der Daten nicht stellen, da deutlich gekennzeichnet ist, dass die Daten rein automatisch erhoben sind und nicht intellektuell nachbearbeitet wurden. Hier scheinen also übliche Maßstäbe – wie sie sich für die (Print-)Lexikographie etabliert haben – nicht angemessen. Deshalb ist es notwendig, terminologische Unterscheidungsmöglichkeiten zu entwickeln, mit denen man deutlich machen kann, dass es sich beim Wortschatz-Lexikon um eine andere Art von elektronischem Nachschlagewerk handelt als beispielsweise bei einer elektronischen Version des „Großen Wörterbuchs zur deutschen Sprache“ aus dem Dudenverlag und dementsprechend auch andere Bewertungsmaßstäbe angelegt werden müssen. Solche begrifflichen Unterscheidungen bieten dabei selbst noch keine Bewertungsmöglichkeiten, sondern sollen mögliche Ordnungsvorstellungen liefern, nach denen dann Bewertungskriterien entwickelt oder schon entwickelte gruppiert werden können. Im Folgenden wird es daher um erweitertes ‚terminologisches Handwerkszeug’ für die Wörterbuchkritik gehen.

Um eine solche Unterscheidung in verschiedene Arten der Erstellung elektronischer (Sprach-)Nachschlagewerke vornehmen zu können, soll zunächst geklärt werden, was in diesem Untersuchungsrahmen unter einem elektronischen Wörterbuch verstanden werden soll und was nicht.

2.2 Was ist ein elektronisches Wörterbuch?

Die Bezeichnung elektronisches Wörterbuch ist – allgemeinsprachlich formuliert – genau genommen nicht passend für den Bezugsgegenstand, da ein elektronischer Datenträger kein *Buch* ist. Die Übertragung des Begriffs Wörterbuch auf die Publikation lexikographischer

Daten auf einem elektronischen Datenträger kann aber dann hilfreich sein, wenn damit auch bestimmte Eigenschaften des gedruckten Wörterbuchs auf das elektronische Wörterbuch übertragen werden können, d.h. wenn damit zur kommunikativ adäquaten Verwendung beigetragen wird. Dies ist dann gegeben, wenn die übertragenen Eigenschaften dem Bezugsgegenstand entsprechen.[5]

Um genauer zu prüfen, inwiefern und auch für welche Art von elektronisch publizierten lexikographischen Daten die Bezeichnung elektronisches Wörterbuch sinnvoll ist, ist zunächst zu fragen, wie ein gedrucktes Sprachwörterbuch charakterisiert werden kann. WIEGAND[6] definiert zunächst ein Nachschlagewerk:

Def. 1: „Ein *Nachschlagewer*k ist ein *Buch* (hier verstanden als etwas Gedrucktes) mit wenigstens einer definierten äußeren Zugriffsstruktur, dessen genuiner Zweck darin besteht, daß ein potentieller Benutzer aus den lexikographischen Textdaten Informationen zum Gegenstandsbereich des Nachschlagewerkes gewinnen kann." (WIEGAND 1998, 58)

Darauf aufbauend definiert WIEGAND ein Sprachwörterbuch:

Def. 2: „Ein *Sprachwörterbuch* ist ein *Nachschlagewerk*, dessen genuiner Zweck darin besteht, daß ein potentieller Benutzer aus den lexikographischen Textdaten Informationen zu sprachlichen Gegenständen gewinnen kann." (WIEGAND 1998, 58)

Folgt man dieser Definition, ist der Terminus elektronisches Wörterbuch demnach nur dann für eine kommunikativ adäquate Verwendung im fachsprachlichen Kontext hilfreich, wenn die Definitionen WIEGANDS für gedruckte Nachschlagewerke und Sprachwörterbücher folgendermaßen auf die Definitionen für elektronische Nachschlagewerke und Wörterbücher übertragen werden können:

Def. 3: Ein *elektronisches Nachschlagewer*k ist eine *elektronisch verfügbare Datensammlung* mit wenigstens einer definierten äußeren Zugriffsstruktur, dessen genuiner Zweck darin besteht, daß ein potentieller Benutzer aus den lexikographischen Textdaten Informationen zum Gegenstandsbereich des elektronischen Nachschlagewerkes gewinnen kann.

Def. 4: Ein *elektronisches Sprachwörterbuch* ist ein *elektronisches Nachschlagewerk*, dessen genuiner Zweck darin besteht, daß ein potentieller Benutzer aus den lexikographischen Textdaten Informationen zu sprachlichen Gegenständen gewinnen kann.

Nun bezeichnet aber WIEGAND z.B. auch Komponenten eines Sprachübersetzungssystems, welche als eine Art Wörterbuch fungieren, als Wörterbuch, genauer als Maschinenwörterbuch (WIEGAND 1998, 241), bei dem es entweder keine Benutzer gibt oder sozusagen der Computer der Benutzer ist. Ist es sinnvoll, auch in diesem Fall von einem Wörterbuch zu

5 Vgl. zu Metaphern im Bereich einer neuen Technologie FREISLER (1994), 33 (Fußnote 37), hier in Bezug auf Sprachmetaphern im Bereich Computer: „WEINGARTEN weist mit Recht darauf hin, dass die enorme Vielfalt der Dialogmetaphern, die in bezug auf den Computer existieren, wohl am besten als Deutungsversuche betrachtet werden können, um eine neue Technologie dem alltäglichen und wissenschaftlichen Verstehenshorizont zugänglich zu machen. Es ist bekannt, dass solche Eingliederungsversuche bzw. die Subsumtion von Unbekanntem unter Vertrautes in einer ersten Phase meist mit Hilfe von metaphorischen Übertragungen vorgenommen werden." Insofern könnte auch die Benennung elektronischer Sprachnachschlagewerke als elektronische Wörterbücher eine Übergangserscheinung sein.

6 Die folgenden Überlegungen gehen im wesentlichen von WIEGANDS *Wörterbuchforschung* (WIEGAND 1998) aus, da die Ausführungen dort m.W. die ausführlichsten und genauesten Überlegungen zu der Frage darstellen, was ein Sprachwörterbuch ist.

sprechen? Der Benutzer in den Definitionen 1–4 ist ein Handelnder, der „Informationen gewinnen" kann; er muss daher ein Aktor mit kognitiven Fähigkeiten sein. Der Computer ist jedoch kein Aktor mit kognitiven Fähigkeiten, sondern eine Maschine, die vom Menschen dazu programmiert werden kann, Daten zu verarbeiten; ein Computer kann aus Daten keine Informationen gewinnen.[7] Davon ausgehend ist es m.E. sinnvoll, elektronische Wörterbücher nur dann als Wörterbücher zu bezeichnen, wenn sie für einen menschlichen Benutzer gemacht sind. Nur dann können grundsätzliche Eigenschaften eines gedruckten Wörterbuchs sinnvoll auf das elektronische Wörterbuch übertragen werden. Ist der Computer als Benutzer gedacht, sollte besser von „lexikalischen Ressourcen für sprachtechnologische Produkte" (ENGELBERG/LEMNITZER 2001, 230)[8] gesprochen werden, z.B.: ein automatisches Übersetzungsprogramm basierend auf einer lexikalischen Ressource. Die Definitionen 1–4 sollten demnach dahingehend ergänzt werden, dass mit „potentiellem Benutzer" immer ein Mensch gemeint ist. Dass damit die Verwendung des Begriffs elektronisches Wörterbuch an die Verwendung des Begriffs Wörterbuch in der printlexikographisch orientierten Wörterbuchforschung angelehnt werden soll, ist dabei nicht primär durch den angestrebten Anschluss an diese Forschungen motiviert. Vielmehr scheint mir diese Einengung des Begriffs deshalb angemessen, weil nur so grundsätzliche Eigenschaften und Kriterien, die Wörterbücher ausmachen und die über die Präsentation in verschiedenen Medien Bestand haben, sinnvoll übertragen werden können.

Nun wurde der Begriff (elektronisches) Wörterbuch dahingehend eingegrenzt, dass als Benutzer immer ein Mensch vorausgesetzt wird; d.h. dass es ein Produkt *für Menschen* ist. Das Beispiel aus dem Wortschatz-Lexikon hatte jedoch mehr zum Gegenstand, inwieweit die Daten *von Menschen* erarbeitet wurden. Letzteres soll in Abschnitt 2.4 in die Unterscheidung von automatisch erstellten vs. lexikographisch, d.h. menschlich bearbeiteten elektronischen Wörterbüchern aufgenommen werden. Zunächst wird jedoch ein Vorschlag zu einer alternativen Benennung zu elektronischem Wörterbuch eingeführt.

2.3 Elektronisches Wörterbuch – Wortschatzinformationssystem

Wie in 2.2 anfangs herausgestellt, ist die Benennung elektronisches Wörterbuch genau genommen nicht passend, da mit dem bezeichneten Bezugsgegenstand kein Buch vorliegt. Im vorhergehenden Abschnitt wurde die Benennung von der Definition her jedoch so eingegrenzt, dass sie zumindest hilfreich für die kommunikativ adäquate Verwendung im fach-

7 vgl. WIEGAND (1998), 171: „Menschen verarbeiten Daten zu Informationen und Informationen, über welche sie bereits verfügen, zu anderen Informationen; sie *erarbeiten* sich Informationen (und damit Kenntnisse von etwas) in Wahrnehmungs- und Denkprozessen, um sich in der Welt orientieren und insbesondere, um zielorientiert handeln zu können. Wahrnehmungs- und Denkprozesse gelten daher als *informationsverarbeitende*, intraindividuelle Prozesse. Computer werden in Handlungszusammenhängen als handlungsunterstützendes Mittel zur Erreichung von Handlungszielen von Menschen derart eingesetzt, dass sie in *datenverarbeitenden* extraindividuellen Prozessen Daten zu anderen Daten *verarbeiten*. Da alle computerinternen Daten als Mengen von numerischen Zeichenexemplaren aufgefasst werden können, sind datenverarbeitende niemals informationsverarbeitende Prozesse."

8 ENGELBERG/LEMNITZER sprechen in diesem Zusammenhang neben der genannten Redeweise auch von „lexikalischen Komponenten sprachverarbeitender Systeme" (ENGELBERG/LEMNITZER 2001, 204).

sprachlichen Kontext sein kann. Trotz dieser Eingrenzung bleibt das o.g. ‚Defizit' bestehen. Aus diesem Grund soll hier eine mögliche alternative Benennung eingeführt werden.

Wie könnte also ein wortschatzbezogenes elektronisches Nachschlagewerk genannt werden, wenn man den Begriff Wörter*buch* darin nicht verwenden will? Mein Vorschlag lautet: als grundsätzliche Bezeichnung für sprachbezogene elektronische Nachschlagewerke soll *Wortschatzinformationssystem* dienen. Diese Benennung hat zwei wesentliche Vorteile: Sie bedient sich nicht der Buchmetapher und die Einordnung als *Informationssystem* kommt von der Bezeichnung her der neuen Art des Arbeitens mit einem elektronischem Sprachnachschlagewerk näher, da sie mehr Dynamik in der Datenabfrage vermuten lässt. Wichtig ist jedoch zu beachten, dass Informationssystem hier in dem Sinne gemeint ist, dass in einem solchen System Daten dargeboten sind, aus denen sich die Benutzer Informationen erschließen können. Durch die Sortierung sind die Daten zu potentieller Information aufbereitet. Das Wortschatzinformationssystem enthält somit keine Informationen, es enthält nur Daten, die zu potentieller Information aufbereitet sind.

Demnach ist zunächst ein Informationssystem zu definieren:

Def. 5: Ein Informationssystem ist eine *elektronisch verfügbare Datensammlung* mit wenigstens einer definierten äußeren Zugriffsstruktur, dessen genuiner Zweck darin besteht, daß ein potentieller menschlicher Benutzer aus den zugreifbaren Daten Informationen zum Gegenstandsbereich des Informationssystems gewinnen kann.

Ein Wortschatzinformationssystem ist darauf aufbauend folgendermaßen zu definieren:

Def. 6: Ein *Wortschatzinformationssystem* ist ein *Informationssystem*, dessen genuiner Zweck darin besteht, daß ein potentieller menschlicher Benutzer aus den zugreifbaren Daten Informationen zu sprachlichen Gegenständen gewinnen kann.

Der Terminus *Wortschatzinformationssystem* kann damit von seiner Definition her in der metalexikographischen Forschung *synonym* mit dem Terminus *elektronisches Wörterbuch* verwendet werden, auch wenn evtl. andere Konnotationen damit verbunden sind. Hier wird im Folgenden – vor allem in diesem Kapitel – meist die Benennung „Wortschatzinformationssystem" verwendet; sie könnte jedoch an allen Stellen durch „elektronisches Wörterbuch" ersetzt werden. Analog zu elektronischen Wörterbüchern können darüber hinaus auch nähere Einordnungen eines Wortschatzinformationssystems vorgenommen werden, z.B. ein „allgemeines, einsprachiges Wortschatzinformationssystem" oder ein „zweisprachiges Wortschatzinformationssystem englisch-deutsch/deutsch-englisch".

2.4 Automatisch erstellte vs. lexikographisch bearbeitete Wortschatzinformationssysteme

Im Ausgangsbeispiel wurde die Frage aufgeworfen, ob und wie man Wortschatzinformationssysteme, deren Daten rein automatisch aus Textkorpora erhoben wurden von solchen Wortschatzinformationssystemen unterscheiden kann, die von Lexikographen erarbeitet wurden. [9] In dem in 2.1. genannten Beispiel spielen die Menschen im Erarbeitungsprozess zwar auch eine wichtige Rolle, aber eine *andere* Rolle als üblicherweise Lexikographinnen:

9 Für diesen Abschnitt danke ich vor allem CYRIL BELICA und KATHRIN STEYER für wertvolle Anregungen.

nämlich v.a. in der Entwicklung der Korpusanalysemethoden und auch in der Zusammenstellung der zugrunde gelegten Korpora. Der Unterschied zwischen rein automatisch erstellten vs. menschlich bearbeiteten Wortschatzinformationsssystemen ist, dass die automatisch gewonnenen Daten nicht von Lexikographen sortiert und bewertet werden. Dabei sind diese Formen der Datenerabeitung keine sich gegenseitig ausschließenden Erstellungsformen. Im Gegenteil: sie können sich aneinander anschließen oder sogar im Zuge der Erarbeitung in einem wechselseitigen Prozess immer wieder angewendet werden. Hier interessiert jedoch vordringlich nicht der Verlauf und die Phasen der automatischen oder automatisch unterstützten Erstellung von Wortschatzinformationssystemen, sondern der Status der Daten in einem Wortschatzinformationssystem, wenn es publiziert, also z.B. im Internet der Öffentlichkeit zugänglich gemacht wird.

Wortschatzinformationssysteme können dabei analog zu ihrer Datenbasis wie in Abbildung 3 gezeigt unterschieden werden.

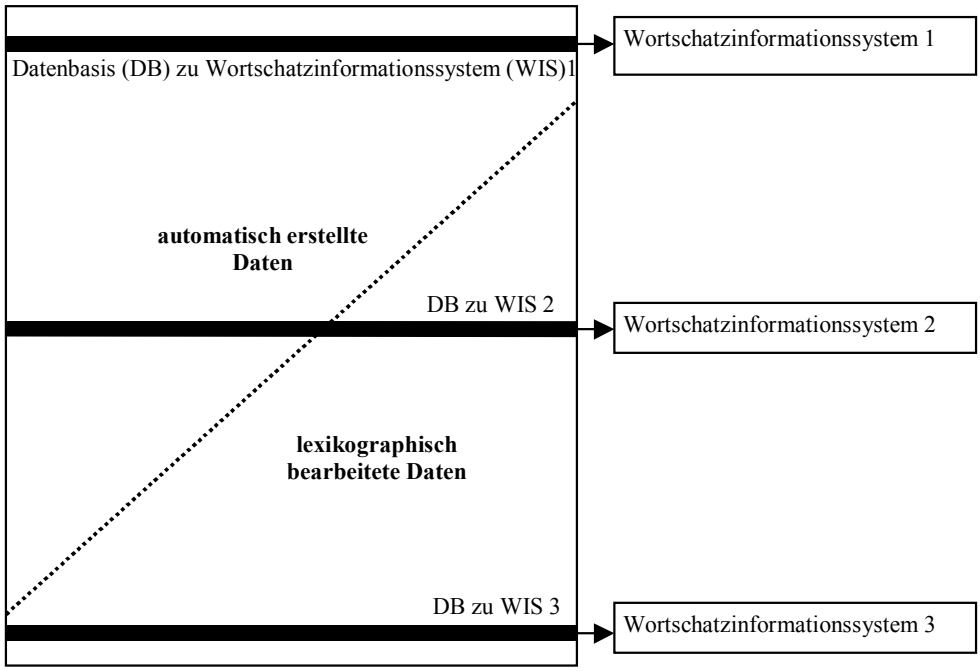

Abb. 3: Arten von Wortschatzinformationssystemen unterschieden nach ihrer Datenbasis
„⟶" bedeutet soviel wie: „aus dieser Datenbasis wird entwickelt"

Die Abbildung zeigt einen fiktiven Datenpool, der aus automatisch erstellten und lexikographisch bearbeiteten Daten besteht. Bei den Daten handelt es sich um Sprachdaten. Aus diesem fiktiven Datenpool wird eine konkrete Datenbasis herausgegriffen, die durch die graue Linie veranschaulicht wird; aus dieser konkreten Datenbasis wird ein Wortschatzinformationssystem entwickelt. Diese Wortschatzinformationssysteme können aufgrund der Art ihrer Datengrundlage unterschieden werden. Das erste Wortschatzinformationssystem besteht nur aus automatisch erstellten Daten, das zweite sowohl aus automatisch erstellten

als auch aus lexikographisch bearbeiteten Daten und das Wortschatzinformationssystem 3 besteht nur aus lexikographisch bearbeiteten Daten. Dieser fiktive Datenpool kann in einem konkreten Projekt von oben nach unten betrachtet den zeitlichen Verlauf eines Projektes darstellen, wenn zunächst eine Menge automatisch erstellter Daten vorliegt, die dann sukzessive lexikographisch bearbeitet wird. Es ist aber auch ein Projekt denkbar, bei dem nur ein Wortschatzinformationssystem der ersten Art entsteht.

Unter lexikographischer Bearbeitung wird hier jede Art der reflektierten menschlichen Bearbeitung der automatisch erstellten Daten verstanden, vom Überprüfen über das Umsortieren bis hin zum Kommentieren. Es ist also mit lexikographischer Bearbeitung nicht gesagt, ob das objektsprachliche Material durch Kommentierungen o.ä. angereichert wird.[10] Die lexikographische Bearbeitung steht allein dafür, dass die Daten von Experten reflektiert gesichtet und überprüft wurden und darüber hinaus eventuell kommentiert sind.

Das WORTSCHATZ-LEXIKON, so wie es oben gezeigt wurde, ist der ersten Art eines Wortschatzinformationssystems zuzuordnen, besteht also nur aus automatisch erstellten Daten. Manche Artikel des WORTSCHATZ-LEXIKONs sind eventuell auch menschlich bearbeitet, allerdings ist dem Benutzer nicht ersichtlich, ob Angaben schon überprüft wurden oder nicht. Nach der einheitlichen Angabe unter den Einträgen im WORTSCHATZ-LEXIKON kann das Produkt als Ganzes nur der ersten Art zugeordnet werden.

Wichtig ist festzuhalten: Diese verschiedenen Arten von Wortschatzinformationssystemen sollen nicht als qualitative Stufen verstanden werden, d.h. ein Nachschlagewerk von der ersten Art ist ein schlechtes und dann wird es Stufe für Stufe besser. Innerhalb jeder dieser Arten kann es große graduelle Unterschiede in der Qualität geben. Sind z.B. bei einem zunächst automatisch erstellten Wortschatzinformationssystem schon das zugrunde gelegte Textkorpus und v.a. die Analysemethode nicht gut, dann kann an den Daten viel bearbeitet werden, ohne dass das Produkt wirklich überzeugend wird. Was hier aber Gegenstand der Betrachtung ist und zu der vorgeschlagenen Aufteilung in drei Arten führt, ist eine Klassifikation nach dem Status der Daten im Wortschatzinformationssystem. Dieser ist in jeder dieser Arten unterschiedlich, egal von welcher Qualität die Daten sind.

Wortschatzinformationssysteme der ersten und zweiten Art können Phasen in einem Herstellungsprozess sein, in dem das Ziel verfolgt wird, ein Wortschatzinformationssystem zu erarbeiten, welches vollständig aus lexikographisch bearbeiteten Daten besteht. WIEGAND sagt – für die Printlexikographie – zur Einordnung von Zwischenprodukten in einem lexikographischen Prozess: „Es ist unbedingt zu beachten, dass erweiterte lexikographische Listen keine Wörterbücher sind, sondern Zwischenprodukte in einem computerunterstützten lexikographischen Prozess, dessen Endprodukt ein gedrucktes Wörterbuch ist.“ (WIEGAND 1998, 199). Und um Zwischenprodukte nicht mit Endprodukten zu verwechseln, sollten Zwischenprodukte nicht mit „-wörterbuch“ oder „-lexikon“ bezeichnet werden, z.B. nicht „Maschinenwörterbuch“ (ebd., 201) für einen automatisch erstellten Index. Im vorliegenden Fall – beim WORTSCHATZ-LEXIKON der Universität Leipzig – ist es aber nicht richtig, dass jetzt im Internet zur Verfügung stehende Produkt als ein Zwischenprodukt anzusehen, denn es ist gar nicht Intention der Projektverantwortlichen, alle Einträge von Men-

10 Allerdings ist die menschliche Überprüfung und eine ggf. vorzunehmende Korrektur in ihrem Wert nicht zu unterschätzen. Vgl. dazu beispielsweise BOLTER (1989), 132: “The computer makes visible what writers have always known: that the identifying and arraging of topics is itself an act of writing.” (Zitiert nach FREISLER 1994, 42)

schen zu überprüfen oder sogar zu bearbeiten. Sobald ein Wortschatzinformationssystem also publiziert wird, ist es als (vorläufiges) Endprodukt anzusehen. Trotzdem erfüllen die dort dargebotenen Einträge nicht die traditionellen Eigenschaften der Lexikographie, die medienübergreifend gelten können, wie:

> „Lexikographie ist nicht nur die vermeintlich objektive Präsentation von sprachlichen Fakten, nicht nur interessenloses Zusammenstellen von Daten, sondern auch interessenverhaftetes Schreiben von Texten, damit geistige Verarbeitung von Daten zu neuen Informationen und damit Selektion." (WIEGAND 1998, 60)

Auch wenn Daten, die mit komplexen Analysemethoden aus Korpora gewonnen werden und automatisch verknüpft werden können, in ihrem Wert nicht mit Listen verglichen werden können, und auch wenn diese Analyse nicht „interessenlos" ist, so ist diese Form der Datenerarbeitung zunächst eine andere und die Menschen spielen darin eine andere Rolle, nämlich bis hin zu der automatischen Erstellung der Daten, nicht in ihrer weiteren Bearbeitung.

Allein die menschliche Bearbeitung der Daten soll daher explizit als lexikographische Bearbeitung gelten. Der gesamte Gegenstandsbereich der Lexikographie und auch der Wörterbuchforschung hat sich mit den neuen Möglichkeiten der computerunterstützten oder computerlexikographischen Erstellung von lexikographischen Produkten allerdings erweitert. Gerade durch die Kostengünstigkeit der elektronischen Publikation von Daten und die Möglichkeit ihrer ständigen Aktualisierung kommen ganz neue lexikographische Produkte auf den Markt. Automatisch erstellte Wortschatzinformationssysteme sind damit auch Gegenstand der Wörterbuchforschung. Dies schon allein deshalb, weil man vermuten kann, dass die Benutzer dieser Gebrauchsgegenstände hier keine genuinen Unterschiede in den Produkten sehen, da der Zweck, weshalb solche Gebrauchsgegenstände erstellt werden, sich auf einer allgemeinen Ebene gleichen: nämlich, dass sie benutzt werden können, um aus den lexikographischen Textdaten Informationen über den jeweiligen Gegenstand des Sprachnachschlagewerkes zu gewinnen.[11] Insofern müssen sie in den Gegenstandsbereich der Wörterbuchkritik einbezogen werden. Die Eigenschaften von Lexikographie als eigenständiger kultureller und wissenschaftlicher Praxis (WIEGAND 1998, 41), so wie sie sich in der Printlexikographie entwickelt haben, sind zu großen Teilen jedoch nur auf lexikographisch bearbeitete Wortschatzinformationssysteme übertragbar. Rein automatisch erstellte Wortschatzinformationssysteme sollten daher deutlich davon abgegrenzt werden. Um es noch einmal herauszustellen: Für automatisch erstellte Wortschatzinformationssysteme bzw. automatisch erstellte elektronische Wörterbücher gilt m.E., dass sie dem Gegenstandsbereich der Lexikographie zuzuordnen sind, da sich die Ergebnisse hinsichtlich ihres

11 Konsequent vom Benutzer auszugehen, wird z.B. in der Funktionslehre von BERGENHOLTZ und TARP (z.B. BERGENHOLTZ/TARP 2002) gefordert. Die Benutzer werden das Wortschatz-Lexikon wie schon gesagt wahrscheinlich nicht prinzipiell anders benutzen als ein anderes Wortschatzinformationssystem, in dem die Daten lexikographisch bearbeitet sind. Für die Wörterbuchforschung scheint eine deutliche Unterscheidung aber dennoch notwendig zu sein. Und auch von Benutzerseite ist diese Unterscheidung zu motivieren: denn auch hier wäre es wichtig, bei jedem Eintrag transparent zu machen, ob es sich um von Menschen überprüfte Daten handelt oder nicht, da die Daten dadurch ein verschiedenes Maß an Verbindlichkeit haben. Außerdem unterscheidet sich die Vorgehensweise und die Rolle der Lexikographen, was auch für die moderne lexikographische Funktionslehre eine Rolle spielt. Ob und wie die hier getroffenen Unterscheidungen in diesem Zusammenhang relevant sind, muss jedoch noch geprüft werden.

grundsätzlichen Herstellungszweckes gleichen; auch ist hier anzuwenden, dass „Lexikographie aus menschlichen Handlungen und ihren Ergebnissen“ (WIEGAND 1998, 52) besteht, denn die Menschen spielen wie gesagt im Erarbeitungsprozess eine Rolle. Die Art dieser menschlichen Handlungen unterscheidet sich jedoch sehr stark von bisherigen lexikographischen Tätigkeiten und demnach unterscheiden sich auch die Maßstäbe, wie die Ergebnisse dieser Handlungen beurteilt werden können. Hier wird demnach die These vertreten, dass sich das Praxisfeld der Lexikographie mit den neuen Möglichkeiten des elektronischen Mediums erweitert hat und damit auch die dazugehörige Forschung einen erweiterten Gegenstandsbereich hat, dass diese Erweiterung jedoch erfordert, dass innerhalb dieses Praxis- und Forschungsfeldes deutliche Unterschiede gezogen werden.

Es ist daher notwendig, den Wortschatzinformationssystemen, deren Daten nicht *von* Menschen bearbeitet wurden, eine Zusatzbezeichnung zu geben, die diesen Unterschied deutlich macht. Dies soll – wie hier schon praktiziert – über Attribute geschehen, die die einzelnen Formen von Wortschatzinformationssystemen spezifizieren. Aufgeteilt in die o.g. Arten sieht das wie in Abbildung 4 gezeigt aus.

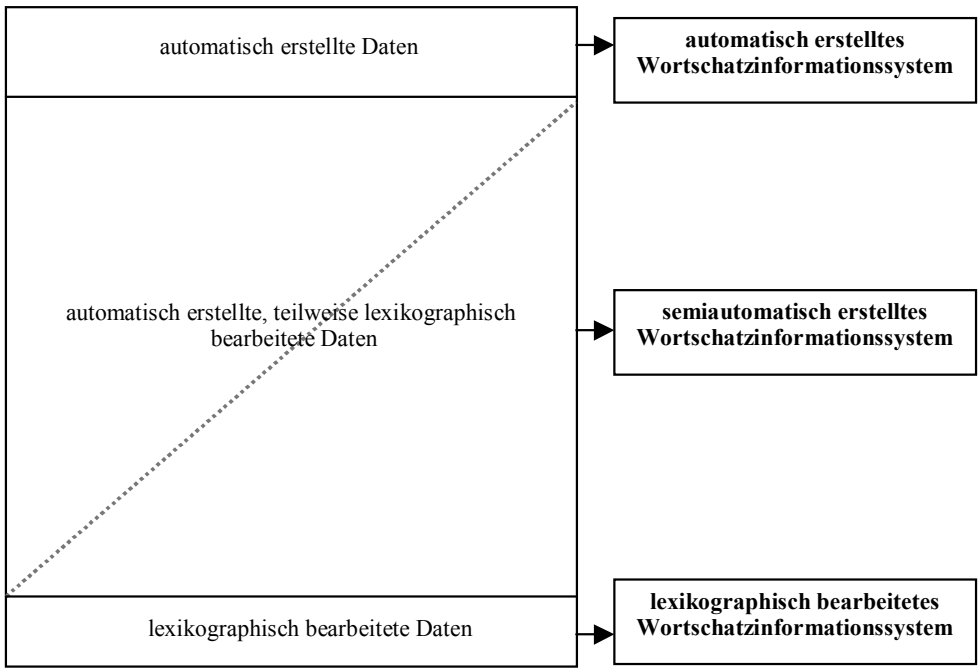

Abb. 4: Arten von Wortschatzinformationssystemen unterschieden nach ihrer Datenbasis
„⟶“ bedeutet soviel wie: „aus einer solchen Datenbasis wird entwickelt“

Diese verschiedenen Arten von Wortschatzinformationssystemen sind daher folgendermaßen zu definieren:

Def. 7: Ein *automatisch erstelltes Wortschatzinformationssystem* ist ein *Wortschatzinformationssystem*, dessen zugreifbare Daten rein automatisch erstellt sind.

Def. 8: Ein *semiautomatisch erstelltes Wortschatzinformationssystem* ist ein *Wortschatzinformationssystem*, dessen zugreifbare Daten automatisch erstellt und teilweise lexikographisch bearbeitet sind.[12]

Def. 9: Ein *lexikographisches Wortschatzinformationssystem* ist ein Wortschatzinformationssystem, dessen zugreifbare Daten lexikographisch bearbeitet sind.

Ist die Datengrundlage eines lexikographisch bearbeiteten Wortschatzinformationssystems zunächst aus elektronischen Korpora automatisch erstellt, so könnte dieses z.B. als „zunächst automatisch erstelltes, vollständig lexikographisch bearbeitetes Wortschatzinformationssystem" näher spezifiziert werden.

Die Unterscheidung in automatisch erstellte vs. lexikographisch bearbeitete Daten ist in den oben entwickelten Unterscheidungen auf der Ebene des gesamten Produktes angesetzt. Bei Wortschatzinformationssystemen, in denen die Daten automatisch erstellt und z.T. menschlich bearbeitet sind, also bei einem semiautomatisch erstellten Wortschatzinformationssystem, muss diese Unterscheidung bei genauer Betrachtung auf einzelne Einträge und vielleicht sogar auf die Ebene einzelner Angaben bezogen werden. Sind immer ganze Einträge menschlich bearbeitet, d.h. immer einzelne Teile der Lemmastrecke rein automatisch erstellt bzw. schon menschlich bearbeitet, dann kann für einen Teil der Lemmastrecke von einem lexikographisch bearbeiteten Wortschatzinformationssystem, bei dem anderen Teil von einem automatisch erstellten Wortschatzinformationssystem gesprochen werden.

Damit sind terminologische Grundunterscheidungen für Arten von Wortschatzinformationssystemen analog zu der Art ihrer Datenbasis getroffen und zugehörige Benennungen präzisiert. Dem Benutzer eines Wortschatzinformationssystems sollte jederzeit transparent sein, welchen Status die Daten haben, die er vor sich hat und wie verlässlich z.B. die Angaben daher sind. Und auch um ein elektronisches Sprachnachschlagewerk zu beurteilen, sollte dies geklärt werden, da – wie oben schon mehrfach erwähnt – für verschiedene Arten von Wortschatzinformationssystemen jeweils andere Beurteilungsmaßstäbe herangezogen werden müssen. Dabei ist die Unterscheidung von Wortschatzinformationssystemen nach der Art ihrer Datenbasis eine von vielen möglichen Klassifizierungen, die für die Lexikographie und Wörterbuchforschung relevant sind.

Die o.g. Unterscheidung in Arten von Wortschatzinformationssystemen ist noch kein Bewertungsmaßstab für die Wörterbuchkritik. Die in Abbildung 3 und 4 dargestellte Verteilung in automatisch erstellte Daten einerseits und lexikographisch bearbeitete Daten andererseits kann jedoch eventuell operationalisiert werden in dem Sinne, dass man folgende These aufstellen kann: Je mehr Daten nicht menschlich bearbeitet sind, desto mehr spezielles Zusatzwissen ist von potentiellen Benutzern entsprechender Wortschatzinformationssysteme gefordert. Denn bei automatisch erstellten Daten muss der potentielle Benutzer selbst in der Lage sein, in einer Benutzungssituation ‚Schrott' auszusortieren, falsch zugeordnete Angaben als solche zu erkennen etc. Veranschaulicht an einem Beispiel: Am Institut für Deutsche Sprache werden im Teilprojekt „Usuelle Wortverbindungen" (s. STEYER

12 Man kann in den meisten Fällen davon ausgehen, dass in semiautomatisch erstellten Wortschatzinformationssystemen die Daten zunächst grundsätzlich automatisch erstellt werden und diese dann im Zuge der lexikographischen Bearbeitung z.T. überprüft, korrigiert und ggf. ergänzt werden und nicht, dass ein Teil der Daten automatisch erstellt und ein anderer Teil der Daten ganz ohne automatische Unterstützung erstellt werden. Wäre letzteres üblicher, wäre es deutlicher von einem *teilweise automatisch erstellten, teilweise lexikographisch erarbeiteten Wortschatzinformationssystem* zu sprechen.

2004)[13] über eine große Lemmastrecke mit Hilfe des Analysewerkzeuges *Statistische Kollokationsanalyse und Clustering* (BELICA 1995) Kookkurrenzangaben automatisch erstellt, teilweise lexikographisch bearbeitet und daraus gewonnene feste usuelle Wortverbindungen in einzelnen Artikeln lexikographisch beschrieben. Für potentielle Benutzer mit sprachwissenschaftlichen und spezieller korpuslinguistischem Hintergrund kann es in einer Benutzungssituation, in der sie sich entweder über die Güte der automatischen Analysemethode informieren oder sich die Daten zur Entwicklung eines eigenen Kollokationsmodells anschauen wollen, sehr wünschenswert sein, mit den rein automatisch gewonnenen, ‚unverfälschten' Daten zum Kookkurrenzverhalten einzelner Lemmata oder Wortformen arbeiten zu können. Hat ein potentieller Benutzer dieses Hintergrundwissen nicht bzw. ein ganz anders geartetes Informationsbedürfnis – möchte sich z.B. ein nichtmuttersprachlicher Benutzer-in-actu schnell darüber informieren, was die Wortverbindung „Tacheles reden" bedeutet –, dann ist es notwendig für ihn, dass die Daten lexikographisch beschrieben sind. Mit den rein automatisch gewonnenen Angaben wäre ihm in diesem Fall nicht geholfen. Ein weiteres Beispiel aus diesem Projekt, an dem demonstriert werden kann, dass nicht nur das eigene Verfassen von Angaben, sondern auch das menschlich reflektierte Kennzeichnen und Einordnen von automatisch gewonnenen Angaben schon einen Wert darstellt, ist Folgendes: Innerhalb des gesamten automatisch gewonnenen Kookkurrenzpotentials zu einem Ausgangswort, d.h. bei allen Einheiten, die sich kohäsiv zum Ausgangswort verhalten, wird in „Usuelle Wortverbindungen" gekennzeichnet, welche Einheiten Basen bzw. Kollokatoren im HAUSMANNschen Sinne sind. (HAUSMANN 2004) Besonders nichtmuttersprachliche Benutzer interessieren sich meist für diese sprachsystematisch gebundenen Einheiten, die sie zum Sprachenlernen elementar benötigen. So können sich solche Benutzer in einer Benutzungssituation, in der sie sich über die Kollokatoren der Basis „Hund" informieren wollen, aus dem gesamten Kookkurrenzpotential zu „Hund", in denen auch solche Partner wie „Auto", „Herr" oder „Pawlow" zu finden sind, die für sie in dieser Benutzungssituation relevanten Partner wie „beißen", „bellen" oder „ausführen" gesondert anzeigen lassen.

Nach der o.g. These können also bestimmte Benutzungssituationen nicht erfolgreich abgeschlossen werden, wenn Daten wie in einem automatisch erstellten Wortschatzinformationssystem den potentiellen Benutzern roh und unbearbeitet zur Verfügung gestellt werden. Ist ein Wortschatzinformationssystem ein automatisch erstelltes Wortschatzinformationssystem, welches nach außen nicht deutlich als solches etikettiert ist und darüber hinaus den potentiellen Benutzern die Daten ohne weitere Erklärung zur Verfügung stellt, so kann das demnach problematisch sein. Ob bzw. inwiefern diese These zutrifft oder inwiefern sie nützlich für die Wörterbuchkritik sein kann, ist jedoch nur in empirischen Tests zu erforschen.

Nun soll – wie in der Einleitung angekündigt – die Perspektive geändert werden und Arten lexikographischer Prozesse in der Gesamtschau Gegenstand der Betrachtung sein.

13 Dieses Projekt ist Teil des Gesamtprojekts *elexiko*; zu näheren Informationen s. www.elexiko.de.

3 Arten von lexikographischen Prozessen

3.1 All in one: Die Duden Ontologie

Anfangs wurde gesagt, dass an dem nun folgenden Beispiel gezeigt werden soll, wie lexikographische Prozesse heute aussehen können. Es bietet sich also an, eine neuartige Form der Wörterbucherstellung anzuführen, die eine besondere Herausforderung bezüglich der Einbindung in eine Klassifikation lexikographischer Prozesse darstellt. Ein solches Beispiel ist das Projekt der „Duden ontology" (ALEXA et.al. 2002).

Die Wörterbücher aus dem Dudenverlag sind sicherlich die bekanntesten in Deutschland. Schon früh hat man im Bibliographischen Institut den Computer an zahlreichen Stellen im lexikographischen Prozess zur Unterstützung eingesetzt. Auch die ersten elektronischen Wörterbücher kamen recht früh auf den Markt. Als vorrangig wurden jedoch – und werden auch eigentlich heute noch – die gedruckten Versionen der Wörterbücher gesehen; v.a. aus Gründen der Wirtschaftlichkeit (vgl. KLOSA 2001, 98). Seit einiger Zeit werden im Dudenverlag jedoch auch sprachtechnologische Produkte, z.B. Rechtschreibkorrekturprogramme entwickelt. Um diese verschiedenen Herstellungsprozesse und v.a. die z.T. gleichen Daten, die in verschiedenen Produkten verwendet werden, besser einheitlich pflegen zu können, will der Verlag einen für die gängige Praxis sehr innovativen Weg beschreiten. Alle Daten für die verschiedenen Wörterbücher sollen perspektivisch in einem Datenpool vorgehalten werden, aus dem die verschiedenen Wörterbücher in gedruckter und elektronischer Form wie auch die sprachtechnologischen Produkte entwickelt werden sollen; diese Datenbasis wird als „Duden ontology" bezeichnet. (ALEXA et.al. 2002) Als Basis aller Produkte dient also eine Datenbasis, in der alle Daten gepflegt werden. Dies ist eine sehr neuartige Form eines lexikographischen Prozesses, da hier aus einer Datenbasis unterschiedliche Wörterbuchtypen – z.B. ein Rechtschreibwörterbuch wie ein Bedeutungswörterbuch – als auch Wörterbücher auf unterschiedlichen Medien wie auch sprachtechnologische Produkte entwickelt werden. Es versteht sich von selbst, dass diese Form eines lexikographischen Prozesses erst mit der weiten Verbreitung und Anwendung des elektronischen Mediums in der Lexikographie denkbar wurde. Von daher ist es spannend, wie man dieses Beispiel in eine Übersicht von Arten lexikographischer Prozesse einbauen kann.

Dafür soll zunächst von der Übersicht zu lexikographischen Prozessen, so wie sie in WIEGANDS Wörterbuchforschung dargestellt ist, ausgegangen werden; denn die dortigen Ausführungen sind meiner Meinung nach die ausführlichsten zu diesem Thema.

3.2 Arten von lexikographischen Prozessen nach H. E. WIEGAND

Bei WIEGAND werden im Rahmen der wissenschaftlichen Lexikographie zunächst lexikographische Prozesse ohne Computereinsatz von denen mit Computereinsatz unterschieden. Die lexikographischen Prozesse mit Computereinsatz werden wiederum unterteilt in computerunterstützte lexikographische Prozesse mit dem Ziel des gedruckten Wörterbuchs und

computerlexikographische Prozesse mit dem Ziel des elektronischen Wörterbuchs. [14] Neben vielen anderen Unterschieden in den letztgenannten Arten von Prozessen führt WIEGAND als wesentlichen Unterschied die unterschiedliche fachliche Ausbildung der beteiligten Personen, die den jeweiligen Prozess hauptsächlich steuern, an:

> „Während im computerunterstützten lexikographischen Prozess neben den Lexikographen ein Informatiker benötigt wird, wird im computerlexikographischen Prozess neben Informatikern und Computerlinguisten ein Lexikograph benötigt bzw. jemand, der lexikographische und metalexikographische Kenntnisse besitzt." (WIEGAND 1998, 244)

Diese Feststellung WIEGANDS ist für ihn der Grund, die beiden Arten von Prozessen dem Gegenstandsbereich verschiedener wissenschaftlicher Disziplinen zuzuordnen:

> „Die Tatsache, dass die beiden Arten von lexikographischen Prozessen jeweils überwiegend von Wissenschaftlern in Gang gehalten werden, welche eine unterschiedliche akademische Ausbildung erfahren haben, ist der erste deutliche Hinweis darauf, dass es angebracht ist, die computerlexikographischen Prozesse nicht zum Gegenstandsbereich der Wörterbuchforschung zu rechnen. […] Die computerlexikographischen Prozesse gehören m.E. daher zu einer anderen, eigenständigen wissenschaftlichen Praxis, der *Computerlexikographie*. Der Metabereich zu dieser Praxis wäre ein Forschungsfeld, welches *Computerlexikologie* heißen könnte." (WIEGAND 1998, 244f.)

Die Meinung WIEGANDS, dass im computerlexikographischen Prozess wenig lexikographische Kompetenz benötigt wird, lässt sich nur dann nachvollziehen, wenn man annimmt, dass die meisten computerlexikographischen Prozesse auf schon vorhandene lexikographische Daten aus einem computerunterstützten lexikographischen Prozess zurückgreifen. Denn wenn in einem computerlexikographischen Prozess lexikographische Daten neu erarbeitet und erstellt würden, wäre dort genauso viel lexikographische Kompetenz nötig wie im computerunterstützten lexikographischen Prozess. Dass der computerunterstützte Prozess i.d.R. als abgeschlossen vorausgesetzt wird, zeigt sich auch daran, dass WIEGAND die Wiederverwendung der Satzbänder eines Printwörterbuchs in einem computerlexikographischen Prozess sowohl für die maschinelle Weiterverarbeitung als auch für die Überführung in ein PC-Wörterbuch als mögliche Nahtstellen betrachtet, an denen die beiden Arten vom lexikographischen Prozessen in fruchtbarer Weise zusammenwirken können. (WIEGAND 1998, 245) In beiden Arten der Weiterverwendung ist die inhaltliche lexikographische Arbeit schon im Zuge der Erstellung des gedruckten Wörterbuchs getan; für die weitere Verwendung werden laut WIEGAND lediglich lexikographische und metalexikographische Kenntnisse benötigt, um die Daten angemessen zu verstehen, damit sie dann adäquat weiterverwendet werden können.

Die strikte Trennung des computerunterstützten lexikographischen Prozesses vom computerlexikographischen Prozess geht also i.d.R. von einem abgeschlossenen computerunterstützten lexikographischen Prozess aus, an den sich ein computerlexikographischer Prozess anschließen kann. Schematisch kann dieser von WIEGAND dargestellte Zusammenhang folgendermaßen veranschaulicht werden:

14 s. WIEGAND (1998), Abbildung 1–43, 242. Zu lexikographischen Prozessen allgemein s. WIEGAND (1998), 38ff. Zur allgemeinen Erläuterung der Herstellung von Wörterbüchern s. auch ENGELBERG/LEMNITZER (2001), 197ff.

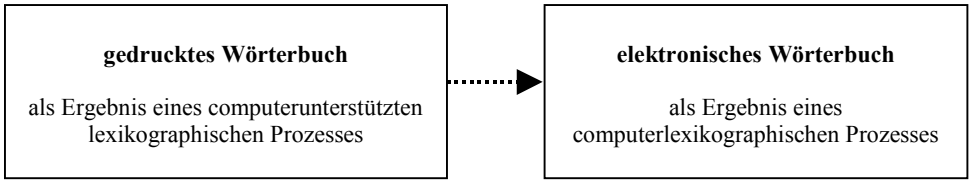

Abb. 5: Veranschaulichung zum möglichen Zusammenwirken computerunterstützter lexikographischer Prozesse und computerlexikographischer Prozesse nach H. E. WIEGAND. „·····▸" bedeutet soviel wie „Ergebnisse werden übernommen für"

Kritisch anzumerken ist: WIEGAND geht davon aus, dass unterschiedliche Publikationsmedien unterschiedliche Herstellungsprozesse bedingen; dies ist auch unbestritten. Diese Herstellungsprozesse werden jedoch darüber hinaus in unterschiedlichen Wissenschaften angesiedelt, obwohl die lexikographischen Daten, d.h. die Inhalte, sich gleichen können. Jedes Medium verlangt, um es sinnvoll einsetzen zu können, bestimmte Kenntnisse. Dass die Eigenschaften des Druckmediums in der Herstellung von Wörterbüchern von den Lexikographen mitgedacht werden, hat sich kulturell herausgebildet, obwohl die Setzerkunst und das gesamte Handwerk der Buchherstellung nicht im Praxisfeld der Lexikographie anzusiedeln sind. Das elektronische Medium ist neu hinzugekommen. Seine Eigenschaften werden daher von Lexikographen heute oft noch nicht selbstverständlich mitgedacht. Dies muss sich erst in Form neuer Prozesse kulturell herausbilden. Dies ist jedoch m.E. noch kein hinreichender Grund dafür, die Arbeiten rund um eine elektronische Publikation lexikographischer Daten in einen anderen Wissenschaftsbereich zu verlegen. Eine solche Rolle verkennt die zentrale Rolle der lexikographischen Daten und damit auch die der Lexikographen. Außerdem haben viele Erfahrungen gezeigt, dass gerade diese strikte Trennung zwischen denjenigen, die Wörterbücher erarbeiten, um sie in gedruckter Form zu publizieren und denjenigen, die diese dann elektronisch weiterverarbeiten, zu einer schlechten Qualität elektronischer Wörterbücher bzw. Wortschatzinformationssystemen[15] führen kann. Die Frage ist außerdem, ob diese scharfe Trennung zwischen computerunterstütztem lexikographischem Prozess und computerlexikographischen Prozess so noch sinnvoll Bestand haben kann, wenn aus *einer* Datenbasis in *einem* Prozess sowohl ein gedrucktes als auch ein elektronisches Wörterbuch entwickelt werden soll, wie es heute schon oft Realität in lexikographischen Projekten ist (s.o.).

Die strikte Trennung von computerunterstütztem lexikographischen Prozess und computerlexikographischem Prozess kommt bei WIEGAND v.a. deshalb zustande, weil er die Eigenschaften des computerlexikographischen Prozessen überwiegend an den von ihm so genannten „Maschinenwörterbüchern" festmacht, also an Wörterbüchern, die den Computer als ‚Benutzer' haben. Unter Ergebnissen von computerlexikographischen Prozessen werden grundsätzlich allerdings auch bei ihm alle Arten der Präsentation lexikographischer Daten auf elektronischen Datenträgern verstanden. Wie unter 2.2 ausgeführt wurde, sollen hier allerdings nur solche Produkte als Wörterbücher bezeichnet werden, die für einen

15 Im Folgenden wird meist die Bezeichnung *elektronisches Wörterbuch* statt *Wortschatzinformationssystem* verwendet, um einen verständlicheren Anschluss an die bisherigen Forschungen über lexikographische Prozesse zu gewährleisten. Es gilt jedoch auch hier wiederum, dass die Bezeichnung *elektronisches Wörterbuch* immer durch *Wortschatzinformationssystem* ersetzt werden könnte.

menschlichen Benutzer entwickelt sind. „Maschinenwörterbücher“ sollen nicht mehr als Wörterbücher klassifiziert werden. In eine erweiterte Übersicht für lexikographische Prozesse muss demnach eingearbeitet werden, dass es Prozesse gibt, in denen elektronische Wörterbücher entstehen mit evtl. gleichzeitig zu entwickelnden lexikalischen Ressourcen sprachtechnologischer Produkte (i.S.v. WIEGANDS Maschinenwörterbüchern). Daneben müssen Prozesse einbezogen werden, in denen gleichzeitig ein gedrucktes und elektronischen Wörterbuch entsteht und eventuell zusätzlich eine lexikalische Ressource für sprachtechnologische Produkte entwickelt wird, so wie es am Beispiel der „Duden ontology“ gezeigt wurde.

3.3 Arten von lexikographischen Prozessen. Eine erweiterte Übersicht

Es bleibt zunächst festzuhalten: Für eine sinnvolle Unterscheidung von lexikographischen Prozessen, auch gerade hinsichtlich ihrer Verankerung in wissenschaftlichen Disziplinen, ist es wichtig zu klären, ob die daraus entstehenden Produkte für einen menschlichen Benutzer entwickelt werden oder nicht. [16] Wenn Wörterbücher für einen menschlichen Benutzer gemacht werden, unterscheiden sich die Handlungen, die zu der Herstellung des Wörterbuchs vollzogen werden, nicht grundsätzlich – egal ob eine gedruckte oder eine elektronische Ausgabe publiziert werden soll. Dass vermehrter interdisziplinärer Austausch nötig ist und dass sich Anforderungen an die beteiligten Personen verändern, wenn elektronische Wörterbücher publiziert werden sollen, ist dabei unbestritten.[17]

Wie oben schon gesagt wurde, muss in eine erweiterte Übersicht lexikographischer Prozesse eine neue Art eingeführt werden, nämlich ein lexikographischer Prozess, in dem gleichzeitig ein gedrucktes und elektronisches Wörterbuch bzw. ein Wortschatzinformationssystem entwickelt wird. Solche lexikographischen Prozesse sind damit *medienneutral konzipiert*. Diese Einordnung soll zunächst näher erläutert werden.

Medienneutralität ist ein Schlagwort, welches immer häufiger verwendet wird in z.T. wenig sinnvollen Verbindungen, wie z.B. als ‚Medienneutrales Publizieren‘. Eine Publikation setzt immer ein Medium als Träger voraus und kann daher nicht medienneutral sein.

16 Mit der Unterscheidung von Mensch oder Computer als Benutzer belegt auch STORRER die o.g. Termini anders als WIEGAND: „Die *computerunterstützte Lexikographie* beschäftigt sich mit den Möglichkeiten, den lexikographischen Arbeitsprozess in all seinen Phasen durch den Computer zu unterstützen. Endprodukt dieses Prozesses sind Wörterbücher für menschliche Benutzer, die als Buch oder mit ‚Neuen Medien‘ vertrieben werden. […] Die *Computerlexikographie* befasst sich mit der Spezifikation lexikalischen Wissens für Systeme der maschinellen Sprachverarbeitung, z.B. Systeme zur Generierung gesprochener Sprache.“ (STORRER 1996, 239) Die andere Verwendung der gleichen Termini scheint jedoch zur Klärung hier nicht günstig.

17 Vgl. dazu auch KLOSA (2001), 100: Sie konstatiert, dass Lexikographen in Zukunft mehr können müssen. „Sie müssen die Bereitschaft mitbringen, in elektronischen Redaktionssystemen die Daten noch konsistenter und penibler einzugeben, als das auf Papier nötig war, damit diese Daten für elektronische Publikationen geeignet sind. Und sie müssen bereit sein, sich gewisse technische und didaktische Kenntnisse anzueignen, damit sie mit Vertreter(inne)n anderer Disziplinen gute CD-ROM-Wörterbücher entwickeln können. Hier zeigt sich, dass sich die Lexikographie insgesamt in einer Phase des Umbruchs befindet. Der Bereich des elektronisches Publizierens entwickelt sich, und damit entwickeln sich auch die Anforderungen an diejenigen, die in diesem Bereich arbeiten.“

Warum wird dieses Schlagwort in der Verbindung „medienneutrale Konzeption“ hier trotzdem verwendet? Die Adjektiv bezeichnet hier – so kurz wie kein anderes Wort oder eine Wortverbindung – eine bestimmte Form der Datenhaltung. Eine Form, die die grundlegende Eigenschaft hat, dass die Daten nicht untrennbar mit den Eigenschaften eines bestimmten Mediums verbunden sind, d.h. nicht an einem Medium ‚kleben’, sondern möglichst unabhängig davon sind. Medienneutrale Datenhaltung heißt damit, dass die Daten so aufbereitet sind, dass aus *einer* Datenbasis Publikationen in *verschiedenen* Medien entwickelt werden können. Dies impliziert, dass nicht *zunächst* eine Publikation in einem Medium abgeschlossen wird und *dann* – in Form einer Zweitverwertung – eine weitere Publikation in einem anderen Medium entwickelt wird, sondern dass zwei Publikation in zwei verschiedenen Medien parallel entwickelt werden. Der Unterschied zum in Abbildung 4 dargestellten Zusammenwirken computerunterstützter und computerlexikographischer Prozesse ist – bezogen aus die Lexikographie – hier, dass aus einer gemeinsamen lexikographischen Datenbasis sowohl ein gedrucktes wie ein elektronisches Wörterbuch entwickelt wird. Dies kann wie in Abbildung 6[18] veranschaulicht werden.

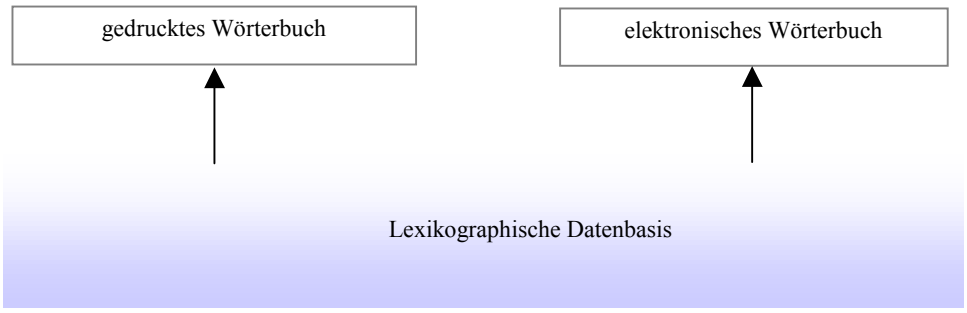

Abb. 6: Veranschaulichung zur Funktion der Datenbasis bei einer gleichzeitigen Entwicklung eines gedruckten und eines elektronischen Wörterbuchs
„⟶“ bedeutet soviel wie: „aus der Datenbasis wird entwickelt“

Mit medienneutraler Datenhaltung ist darüber hinaus meist der Anspruch verbunden, dass die Form der Datenaufbereitung möglichst auch bei künftigen Veränderungen in der Medienlandschaft seine Gültigkeit behält. Sie darf sich also nicht an einzelnen Satzprogrammen, an speziellen elektronischen Datenträgern oder Publikationsmöglichkeiten wie CD-ROM oder Internet orientieren. Die medienneutrale Datenaufbereitung muss sich hingegen an einer Konstante orientieren, die unabhängig von den verschiedenen Medien Gültigkeit besitzt. Diese Konstante kann z.B. der inhaltliche Gehalt der Daten sein. Dies soll an dieser Stelle jedoch nicht weiter ausgeführt werden.

Aufbauend auf 3.2 kann die Gliederung lexikographischer Prozesse nach WIEGAND nun erweitert werden: Innerhalb der lexikographischen Prozesse mit Computereinsatz treten neben die computerunterstützten lexikographischen Prozesse mit dem Ziel des gedruckten Wörterbuchs und die computerlexikographischen Prozesse mit dem Ziel der Publikation oder Verarbeitung lexikographischer Daten auf elektronischen Datenträgern die medienneut-

18 Dies stellt das grundlegende Prinzip eines medienneutral konzipierten lexikographischen Prozesses dar. Genauer s. u. Abbildung 8 in Kapitel 4.

ral konzipierten lexikographischen Prozesse mit dem Ziel des gedruckten *und* elektronischen Wörterbuchs. Außerdem ist die gleichzeitige Verarbeitung der lexikographischen Daten als lexikalische Ressource in sprachtechnologischen Produkten zu integrieren.

Nun ist es allerdings eine Illusion anzunehmen, Daten könnten völlig losgelöst von den Traditionen, die mit den Eigenschaften eines Mediums zusammenhängen, erarbeitet werden. Auch hier sind die Duden-Wörterbücher, so wie sie bisher entwickelt wurden, ein Beispiel: „Altough the majoritiy of the Duden dictionary data are in SGML format, the markup of each dictionary is strongly print orientied rather than content oriented." (ALEXA et. al. 2002, 1). In manchen Stellungnahmen klingt es allerdings, als sei dies ein triviales Problem; beispielsweise schreibt PETELENZ zu Ausgaben in verschiedenen Medien, die auf einer Datengrundlage aufbauen: „Derzeit noch vorrangiges Ziel ist jedoch oft eine Produktionsumgebung, die gleichermaßen die Herstellung einer Printversion ermöglicht. Liegen die Daten in der von mir skizzierten hochstrukturierten Form vor, ist dies ein untergeordnetes Problem, da Standard-Generatoren für die marktüblichen Paper-Engines bereits vorhanden sind." (PETELENZ 1999, 59) So einfach ist die Sache jedoch in der Regel nicht. Die medienneutral konzipierten lexikographischen Prozesse werden daher nochmals unterschieden in diejenigen, in denen vorrangig das gedruckte Wörterbuch geplant wurde und die, in denen vorrangig das elektronische Wörterbuch geplant wurde. Diese Unterscheidung innerhalb der medienneutral konzipierten lexikographischen Prozesse betrifft die Form und Inhalte der lexikographischen Daten, sollte jedoch nicht das Konzept der Datenaufbereitung beeinflussen. Dieses Konzept sollte deshalb immer noch klar unterschieden sein von der Datenaufbereitung lexikographischer Daten, die ausschließlich für die Publikation als gedrucktes oder als elektronisches Wörterbuch konzipiert sind.[19]

19 Es ist zum jetzigen Zeitpunkt noch nicht zu leisten, den medienneutral konzipierten lexikographischen Prozess schon genau zu charakterisieren, wie WIEGAND es fordert: „Es hängt viel davon ab, dass man die spezifisch lexikographischen Tätigkeiten genau charakterisiert, denn nur dann wird Sprachlexikographie lehrbar." (WIEGAND 1998, 47). Um das leisten zu können, muss diese Form von lexikographischem Prozess zunächst in vielen Anwendungen in der Praxis näher untersucht werden.

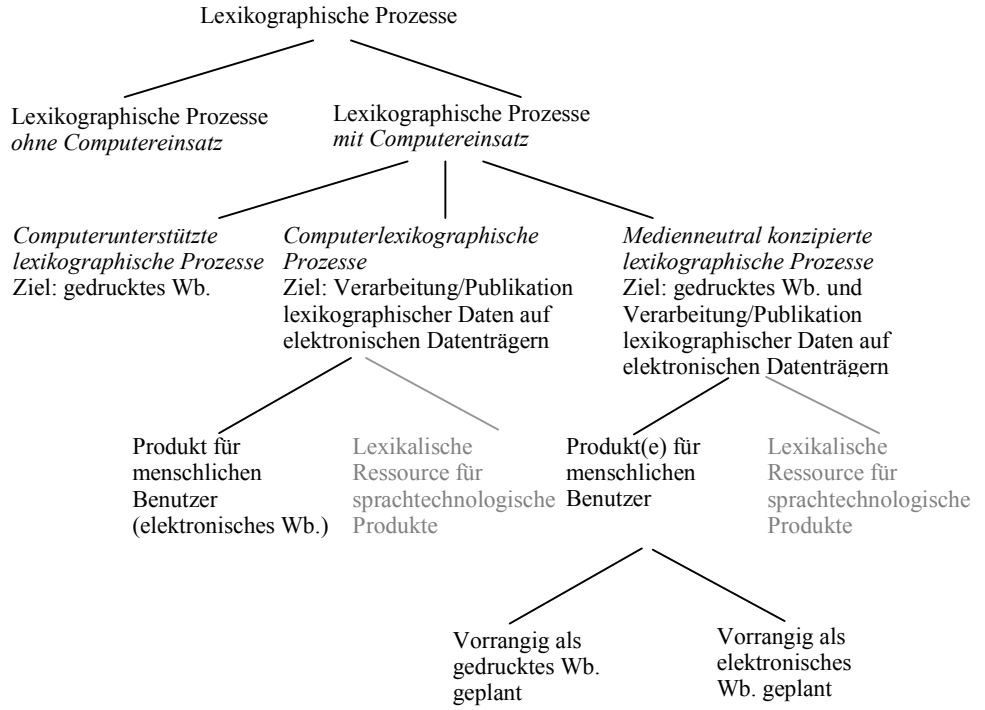

Abb. 7: Erweiterte Übersicht zu Arten von lexikographischen Prozessen
„——" bedeutet soviel wie: „bilden ein Unterklasse von/sind eine Oberklasse zu"

Die erweiterte Übersicht zu Arten lexikographischer Prozesse ist in Abbildung 7 gezeigt. Hierin wurde es aus Platzgründen unterlassen, überall zum „elektronischen Wörterbuch" das „Wortschatzinformationssystem"[20] zu ergänzen. Außerdem handelt es sich an den Stellen, an denen in der Abbildung „Ziel: gedrucktes Wörterbuch" steht, um eine verkürzte Redeweise. Ausführlich und korrekt müsste es heißen: „Ziel: lexikographische Datenbasis, die als gedrucktes Wörterbuch publiziert wird". Dies gilt auch für alle anderen oben dargestellten Arten lexikographischer Prozesse. Als Ergebnis eines lexikographischen Prozesses wird hier also eine lexikographische Datenbasis vorausgesetzt. Warum diese Datenbasis vom publizierten Wörterbuch unterschieden wird, ist Thema des vierten Kapitels.

In der Forschungsliteratur werden oft andere Begriffe als lexikographische Datenbasis verwendet. ENGELBERG und LEMNITZER sprechen beim Ergebnis eines lexikographischen Prozesses z.B. von einer „lexikalischer Datenbank" (ENGELBERG/LEMNITZER 2001, 230) und auch von einer „lexikalischen Ressource" (ebd., 197).[21] M. E. wird oft nicht (mehr) von

20 Bei medienneutral konzipierten lexikographischen Prozessen wird es sich bei dem Ergebnisprodukt meist um ein lexikographisch bearbeitetes Wortschatzinformationssystem handeln. Bei computerlexikographischen Prozessen können alle Arten von Wortschatzinformationssystemen das Ergebnis sein.

21 Sie sprechen allerdings z.T. auch von lexikographischen Datenbasen. So wie hier: „Zum einen werden mit der stärkeren Verbreitung des Mediums und damit steigenden Marktanteilen dieser Produkte lexikographische Datenbasen entstehen, die sowohl in eine Präsentation im Printmedi-

lexikographischen Daten gesprochen, um zu betonen, dass die Daten nicht in der Form wie früher für Printwörterbücher vorliegen oder vorliegen sollten, und dass die Datenaufbereitung sich vielmehr am Wörterbuchgegenstand ausrichtet oder ausrichten soll. Um dies, und auch um die Vernetzung der Daten und die verschiedenen Zugriffsmöglichkeiten zu betonen, wird daher oft von lexikalischen Daten, lexikologischer Strukturierung etc. gesprochen. So z.B. KNOWLES (1987), 25: „Der maschinenlesbare Text eines Wörterbuchs ist bei weitem keine lexikalische Datenbank, da die innere Struktur und das Netzwerk fehlen." In diese Richtung auch STORRER (2001), 61: die Beschreibung linguistischer Merkmale erfordere „ein höheres Maß an Formalisierung als die Erfassung entsprechender Merkmale im gedruckten Wörterbuch"; von daher könne man bei „umsichtiger Modellierung" direkt eine lexikalische Datenbank aufbauen, aus denen dann Produkte für den menschlichen Benutzer genauso „herausgegriffen und in jeweils adäquater Form präsentiert werden" können wie Anwendungen für die maschinelle Sprachverarbeitung. Oder auch GLONING/WELTER (2001), 180: „Das elektronische Wörterbuch als komplex strukturierte lexikologische Datenbasis kann sehr viel mehr sein: ein Abbild der komplexen Zusammenhänge im Wortschatz und ein Informationssystem für sehr unterschiedliche Benutzerinteressen."

Was hier für die Datenaufbereitung in computerlexikographischen und medienneutral konzipierten lexikographischen Prozessen empfohlen wird, ist sicherlich richtig (s. z.B. auch SCHMIDT/MÜLLER 2001). Meiner Ansicht nach spricht jedoch nichts dagegen – trotz einer Datenstrukturierung, die sich nicht an den Gegebenheiten eines einzelnen Mediums orientiert – von einer lexikographischen Datenbasis zu sprechen. Denn was in einer solchen Datenbasis meist Gegenstand der Datenmodellierung ist, ist der genuine Zweck der lexikographischen Angaben. Dieser ist unabhängig vom Publikationsmedium und ist in dieser expliziten Strukturierung auch hilfreich für die Entwicklung sprachtechnologischer Produkte. Der Unterschied zwischen primär rein lexikalischen Ressourcen und lexikographischen Datenbasen scheint mir deshalb noch nicht aufgehoben. Für das zweistellige Prädikat „lexikographische Datenbasis" sprechen daher im Rahmen dieses Themas verschiedene Gründe: Das zweite Glied, die „Datenbasis", ist eine passende Bezeichnung, da sie unspezifisch ist, was die Form der Datenhaltung betrifft. Die Bezeichnung „Datenbank" bezieht dagegen eine bestimmte Form der Speicherung mit ein. Das erste Glied des zweistelligen Prädikats – „lexikographisch" – betont erstens, dass es um lexikographische Prozesse geht, deren Ziel lexikographische Produkte sind und ermöglicht zweitens die Abgrenzung dieser Prozesse gegen solche Projekte wie GermaNet oder WordNet, die als Ziel eine lexikalisch-semantische Datenbasis v.a. für sprachtechnologische Anwendungen entwickeln (KUNZE/WAGNER 2001). Diese terminologische Entscheidung hat nichts damit zu tun, dass die lexikographische Datenbasis vielfältige lexikologisch motivierte Auszeichnungen enthält. Das Ziel soll nur klar herausgestellt werden: primäres Ziel eines lexikographischen Prozesses ist es, Wörterbücher bzw. Wortschatzinformationssysteme als Gebrauchsgegenstände herzustellen. Aus solchen lexikographische Datenbasen können – wie in Abbildung 7 gezeigt – auch lexikalische Ressourcen für die Anwendung in der natürlichen Sprachverarbeitung abgeleitet werden; dies ist jedoch nicht primäres Ziel lexikographischer Prozesse (s. auch das Beispiel der „Duden ontology"). Für das, was aus der lexikographischen Datenbasis für

um als auch für eine Präsentation im elektronischen Medium, unter jeweiliger Berücksichtigung der Möglichkeiten und Grenzen beider Medien, geeignet sind." (ENGELBERG/LEMNITZER 2001, 194)

sprachtechnologische Produkte abgeleitet wird, ist der Begriff der lexikalischen Ressource passend. Diese Unterscheidung zwischen primär rein lexikalischen Ressourcen und lexikographischen Datenbasen ist auch wichtig, um die Berührungspunkte zwischen Lexikographie und natürlicher Sprachverarbeitung klarer sehen zu können. Lexikographische Daten können dann gut in sprachtechnologischen Produkten verwendet werden, wenn ihre Angaben explizit strukturiert sind und so lexikalische Ressourcen daraus abgeleitet werden können. Das Zusammenwirken kann in dieser Weise sehr fruchtbar sein.[22]

Insofern ist hier die Intention entscheidend, ob etwas als lexikographischer Prozess zu bezeichnen ist oder nicht. Ein Herstellungsprozess ist nur dann als lexikographischer Prozess zu bezeichnen, wenn das primäre Ziel die Erarbeitung einer lexikographischen Datenbasis ist, die als gedrucktes und/oder elektronisches Wörterbuch bzw. als Wortschatzinformationssystem publiziert wird. Die Weiterverwendung der lexikographischen Datenbasis als lexikalische Ressource in sprachtechnologischen Produkten ist dabei ein zusätzliche Möglichkeit.

Die Entwicklung lexikalischer Ressourcen bzw. insgesamt die Entwicklung sprachtechnologischer Produkte ist dabei nicht Gegenstand der Sprachlexikographie und der Wörterbuchforschung: hier trifft das zu, was WIEGAND auf den gesamten computerlexikographischen Prozess bezogen hat, nämlich dass die Anforderungen an beteiligte Personen in einem anderen Fachbereich liegen. Dieser ‚Zweig' ist in Abbildung 6 daher in grau dargestellt. Trotzdem ist natürlich eine konstruktive Zusammenarbeit zwischen diesen beiden Bereichen – wie bereits herausgestellt – sehr wichtig und hilfreich.

Nun kann man sich fragen, inwiefern diese Übersicht über Arten lexikographischer Prozesse unmittelbar für die Wörterbuchkritik relevant ist. Denn diese Übersicht dient nicht direkt der Einordnung z.B. verschiedener Arten von Wortschatzinformationssystemen, sondern hat allein zum Gegenstand, wie man lexikographische Herstellungsprozesse in einer Gesamtschau einordnen kann. Diese Einordnung kann jedoch trotzdem hilfreich für die Wörterbuchkritik sein. Ist ein Wörterbuch bzw. ein Wortschatzinformationssystem zu rezensieren, kann es zunächst einer Art von lexikographischem Prozess zugeordnet werden. An ein Wortschatzinformationssystem, welches Ergebnis eines computerlexikographischen Prozesses ist, sind z.B. andere Erwartungen zu richten als an ein Wortschatzinformationssystem, welches das Ergebnis eines medienneutral konzipierten lexikographischen Prozesses ist, in dem vorrangig das Buch geplant wurde. Bei letzterem restringiert schon die Art des Herstellungsprozesses die Möglichkeiten. So können also auch entlang dieser Einordnung schon entwickelte oder noch zu entwickelnde Kriterien zur Bewertung von Wortschatzinformationssystemen spezifisch zugeordnet werden.

Wie in der Einleitung angekündigt, beschäftigt sich der nun folgende dritte Bereich mit einem bestimmten Aspekt der Betrachtung von medienneutral konzipierten und computerlexikographischen Prozessen: eine Sicht auf Ebenen im lexikographischen Prozess.

22 Vgl. KILGARIFF (2000), 105: "NLP needs dictionaries, and dictionary-makers can use NLP to make better dictionaries, so there is a great potential for synergy between the two activities." Weiter ebd., 110: "My purpose in saying this is not to put fear of redundancy in the hearts of lexicographers but to indicate how much more satisfactory their work will become when the tools at their disposal are so much powerful. The techniques tend to find many plausible hypotheses for how a word behaves in a corpus, but are unable to sort the weat from the the chaff, or evidently, to assign meanings to the patterns they find. The lexicographer's task is as before but with less drudgery."

4 Ebenen im lexikographischen Prozess

Zur Eröffnung der neuen Lexicographica-Rubrik „Electronic Dictionaries“ wurde von LEHR die sinnvolle und für Rezensionen schon mehrfach angewendete Unterscheidung zwischen papierorientierten vs. innovativ gestalteten elektronischen Wörterbüchern getroffen:

> „In (meta-)lexikographischer Hinsicht müssen wir zwischen elektronischen Wörterbüchern, die auf ein Papierwörterbuch zurückgehen und solchen, die Neuentwicklungen sind, unterscheiden. Erstere lassen sich außerdem danach subklassifizieren, ob sie eine wesentliche Veränderung bezüglich der Erscheinungsform ihrer Wörterbuchartikel erfahren haben oder nicht, und letztere danach, ob bei der Gestaltung der Wörterbuchartikel an traditionelle lexikographische Formen angeknüpft oder ob ein neuer Weg beschritten wurde – wir sprechen beide Male von *papierorientierten* vs. *innovativen* elektronischen Wörterbüchern.“ (LEHR 1996, 314)

Nun sei ein fiktives Beispiel eines elektronischen Wörterbuchs gegeben, auf das diese Unterscheidung angewendet werden solle. Dieses elektronische Wörterbuch sei Ergebnis eines medienneutral konzipierten lexikographischen Prozesses, in dem vorrangig das Buch geplant wurde. Die Daten in den Wörterbuchartikeln im elektronischen Wörterbuch gleichen weitgehend denen im gedruckten Wörterbuch: Verdichtungen sind zum größten Teil nicht aufgelöst, Ausspracheangaben nach wie vor in Lautschrift und nicht als Tondatei u.v.m. Die Daten scheinen also weitgehend mit denen im gedruckten Wörterbuch übereinzustimmen. Das elektronische Wörterbuch ist daher als papierorientiert zu bezeichnen. Diese Einordnung scheint in anderer Hinsicht jedoch nicht angemessen, denn in der elektronischen Version des Wörterbuchs sind die Zugriffsmöglichkeiten auf die Wörterbuchartikel sehr vielfältig. Es kann gezielt in einzelnen Angaben des Wörterbuchartikels gesucht werden, beispielsweise nur in der Bedeutungsparaphrasenangabe, einzelne Teile der Mikrostruktur können vom Benutzer selektiv zur Anzeige ausgewählt werden etc. In dieser Hinsicht scheint das elektronische Wörterbuch also innovativ gestaltet, da es die Möglichkeiten des neuen Mediums nutzt.

Dieses Beispiel soll demonstrieren, dass man zwischen der Ebene der Datengrundlage eines einzelnen Produkts vs. der Präsentationsebene dieses Produkts, nämlich dem eigentlichen Wörterbuch, unterscheiden sollte, so wie es in Abbildung 8 veranschaulicht wird.

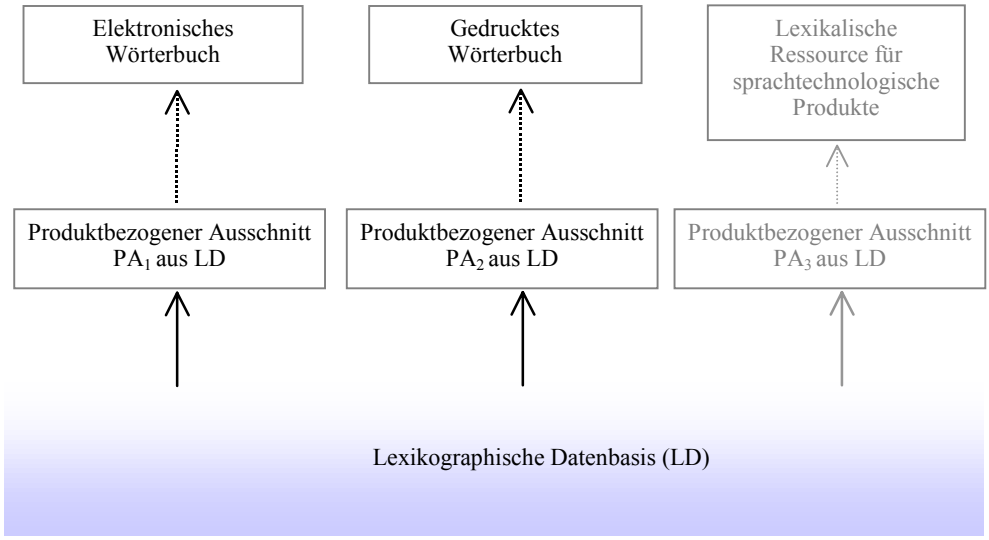

Abb. 8: Ebene der Datenbasis, des produktbezogenen Ausschnitts und des Produkts im lexikographischen Prozess
„⟶" bedeutet soviel wie: „als Ausschnitt wird definiert"
„⋯⋯>" bedeutet soviel wie: „als Produkt wird entwickelt"

In Abbildung 8[23] werden die Ebene der Datenbasis, des produktbezogenen Ausschnitts und des Produkts an einem medienneutral konzipierten lexikographischen Prozess verdeutlicht, aus dem gleichzeitig eine lexikalische Ressource für sprachtechnologische Produkte abgeleitet wird. Letzteres ist aus o.g. Gründen auch in dieser Abbildung grau dargestellt. Diese Unterscheidung der Ebenen des produktbezogenen Ausschnitts und der Präsentation der Daten im Wörterbuch ist deshalb hilfreich, weil sie der Wörterbuchkritik verschiedene Analyseebenen bereitstellt. Denn die Wörterbuchkritik kann so auf zwei Ebenen ansetzen: Auf der ersten Analyseebene können die produktbezogenen Ausschnitte PA_1 und PA_2 verglichen werden, d.h. welche Daten in den beiden Arten von Wörterbüchern enthalten sind. Folgende Vergleiche können z.B. vorgenommen werden:

- Ist PA_1 gleich PA_2?
- Ist PA_1 eine Teilmenge von PA_2?
- Wenn PA_1 eine Teilmenge von PA_2 ist, welche Daten sind dann zusätzlich in PA_2 enthalten?

Zum Beispiel könnte für das oben dargestellte fiktive Beispiel festgestellt werden, dass die produktbezogenen Ausschnitte, also die Datengrundlage für beide Versionen des Wörterbuchs sich gleichen, also $PA_1 = PA_2$ ist.

Die zweite Analyseebene ist die der Präsentation der beiden Wörterbücher. Hier kann die Umsetzung der Daten in die Präsentation beurteilt werden, z.B.: Wie sind die Zugriffsmöglichkeiten auf die Daten? Können Angaben selektiv angezeigt werden? Kann der Be-

23 Zu dem grundsätzlichen Modell in einem anderen Projektzusammenhang s. auch SCHMIDT/MÜLLER (2000).

nutzer selber Notizen machen? etc. So kann u.a. die von LEHR entwickelte Unterscheidung von papierorientiert vs. innovativ gestaltet getrennt auf beide Ebenen bezogen werden.

Diese Unterscheidung der Ebenen der Datenbasis, des produktbezogenen Ausschnitts und des Produkts ist nicht nur für den medienneutral konzipierten lexikographischen Prozess relevant, sondern auch für den computerlexikographischen Prozess. Denn die produktbezogenen Ausschnitte können z.B. auch die Datengrundlagen von zwei verschiedenen elektronischen Wörterbüchern darstellen.

5 Schlussbemerkung

Wie anfangs schon gesagt, sollten in diesem Aufsatz ordnende Betrachtungen zu Wortschatzinformationssystemen bzw. elektronischen Wörterbüchern und lexikographischen Prozessen vorgenommen werden, die hilfreich für die Wörterbuchkritik sein können. Die drei hier ausgeführten Themenbereiche haben dabei einen unterschiedlichen breiten Gegenstand der Betrachtung: Kapitel 2 richtete sich besonders auf die automatisch unterstützte Erstellung von Wortschatzinformationssystemen und beinhaltete terminologische Vorschlägen und dazugehörige Definitionen, wie man automatisch erstellte Wortschatzinformationssysteme von lexikographisch bearbeiteten abgrenzen kann. Der Betrachtungsgegenstand in Kapitel 3 war die Gesamtheit lexikographischer Prozesse und ihre Darstellung in einer der aktuellen Wörterbuchlandschaft angemessenen Übersicht. Im vorangegangenen Kapitel 4 wurde der Fokus wiederum enger auf medienneutral konzipierte lexikographische Prozesse sowie auf computerlexikographische Prozesse gerichtet und dabei eine mögliche Unterscheidung von Analyseebenen entwickelt. Analog zu dieser unterschiedlichen Breite des Betrachtungsgegenstandes sind die Themenbereiche unterschiedlich relevant für die Einordnung einzelner Wortschatzinformationssysteme.

Die Praxis muss nun zeigen, inwiefern die hier getroffenen Unterscheidungen für die Gruppierung von Kriterienkatalogen zur Bewertung verschiedener Arten von Wortschatzinformationssystemen bzw. elektronischer Wörterbücher fruchtbar angewendet werden können.

6 Literatur

ALEXA et. al. 2002 = MELINA ALEXA/BERND KREISSIG/MARTINA LIEPERT/KLAUS REICHENBERGER/KARIN RAUTMANN/WERNER SCHOLZE-STUBENRECHT/SABINE STOYE: The Duden ontology. An Integrated Representation of Lexical and Ontological Information. In: Workshop at I-REC2002. Las Palmas, Gran Canaria (27.5.2002).
Zitiert nach <www.darmstadt.gmd.de/~rostek/alexa-etal-Irec2002.pdf>. 8 Seiten.

BELICA 1995 = CYRIL BELICA: Statistische Kollokationsanalyse und Clustering – COSMAS Analysemodul. © 1995 Institut für Deutsche Sprache. Mannheim.

BERGENHOLTZ/TARP 2002 = HENNING BERGENHOLTZ/SVEN TARP: Die moderne lexikographische Funktionslehre. Diskussionsbeitrag zu neuen und alten Paradigmen, die Wörterbücher als Gebrauchsgegenstände verstehen. In: Lexicographica 18 (2002), 253–263.

DUDEN-DUW3 = DUDEN. DEUTSCHES UNIVERSALWÖRTERBUCH. Mannheim 1996 (3. Aufl.).

ENGELBERG/LEMNITZER 2001 = STEFAN ENGELBERG/LOTHAR LEMNITZER: Lexikographie und Wörterbuchbenutzung. Tübingen 2001 (= Stauffenburg Einführungen ; Bd. 14).

FELDWEG 1997 = HELMUT FELDWEG: Wörterbücher und neue Medien: Alter Wein in neuen Schläuchen? In: Zeitschrift für Literaturwissenschaft und Linguistik 107 (1997), 110–123.

FREISLER 1994 = STEFAN FREISLER: Hypertext – Eine Begriffsbestimmung. In: Deutsche Sprache 1 (1994), 19–50.

GLONING/WELTER 2001 = THOMAS GLONING/RÜDIGER WELTER: Wortschatzarchitektur und elektronische Wörterbücher: Goethes Wortschatz und das Goethe Wörterbuch. In: INGRID LEMBERG/BERNHARD SCHRÖDER/ANGELIKA STORRER (Hg.): Chancen und Perspektiven computergestützter Lexikographie. Tübingen 2001 (Lexicographica. Series Maior 107), 117–132.

HAUSMANN 2004 = FRANZ JOSEF HAUSMANN: Was sind eigentlich Kollokationen? Oder: Wie pervers ist der wissenschaftliche Diskurs? In: KATHRIN STEYER (Hg.): „Den Nagel auf den Kopf treffen". Wortverbindungen mehr oder weniger fest. Berlin, New York 2004 (Jahrbücher des Instituts für Deutsche Sprache 2003) (erscheint).

KILGARIFF 2000 = ADAM KILGARIFF: Business Models for Dictionaries and NLP. In: International Journal of Lexicography 13 (2/2000), 107–118.

KLOSA 2001 = ANNETTE KLOSA: Qualitätskriterien der CD-ROM-Publikation von Wörterbüchern. In: INGRID LEMBERG/BERNHARD SCHRÖDER/ANGELIKA STORRER (Hg.): Chancen und Perspektiven computergestützter Lexikographie. Tübingen 2001 (Lexicographica. Series Maior 107), 93–101.

KNOWLES 1987 = FRANCIS KNOWLES: Möglichkeiten des Computereinsatzes in der Sprachlexikographie. In: HERBERT ERNST WIEGAND (Hg.): Theorie und Praxis des lexikographischen Prozesses bei historischen Wörterbüchern. Akten der Internationalen Fachkonferenz Heidelberg 3.6.–5.6.86. Tübingen 1987 (Lexicographica. Series Maior 23), 11–33.

KUNZE/WAGNER 2001 = CLAUDIA KUNZE/ANDREAS WAGNER: Anwendungsperspektiven des GermaNet, eines lexikalisch-semantischen Netzes für das Deutsche. In: INGRID LEMBERG/BERNHARD SCHRÖDER/ANGELIKA STORRER (Hg.): Chancen und Perspektiven computergestützter Lexikographie. Tübingen 2001 (Lexicographica. Series Maior 107), 229–246.

LEHR 1996 = ANDREA LEHR: Zur neuen Lexicographica-Rubrik „Electronic Dictionaries. In: Lexicographica 12 (1996), 310–317.

PETELENZ 1999 = KRZYSZTOF PETELENZ: Objektorientierte Hypertexterstellung für bilinguale Nachschlagewerke. In: Sprache und Datenverarbeitung 23 (2/1999), 35–62.

QUASTHOFF/WOLFF 1999 = UWE QUASTHOFF/CHRISTIAN WOLFF: Korpuslinguistik und große einsprachige Wörterbücher. In Linguistik online 3 (2/1999)
<www.linguistik-online.de/2_99/quasthoff.html>. 7 Seiten.

SCHMIDT/MÜLLER 2000 = INGRID SCHMIDT/CAROLIN MÜLLER: Zaubernetz. Inhaltsstrukturen und Topic Maps als Potenzial neuer Informationstechnik. In: iX 11 (2000), 100–107.
S. auch <www.heise.de/ix/artikel/2000/11/100/>.

SCHMIDT/MÜLLER 2001 = INGRID SCHMIDT/CAROLIN MÜLLER: Entwicklung eines lexikographischen Modells: Ein neuer Ansatz. In: INGRID LEMBERG/BERNHARD SCHRÖDER/ANGELIKA STORRER (Hg.): Chancen und Perspektiven computergestützter Lexikographie. Tübingen 2001 (Lexicographica. Series Maior 107), 29–52.

STEYER 2004 = KATHRIN STEYER: Kookkurrenz. Linguistisches Modell, Korpusmethodik, lexikografische Perspektiven. In: KATHRIN STEYER (Hg.): „Den Nagel auf den Kopf treffen". Wortverbindungen mehr oder weniger fest. Berlin, New York 2004 (Jahrbücher des Instituts für Deutsche Sprache 2003) (erscheint).

STORRER 1996 = ANGELIKA STORRER: Metalexikographische Methoden in der Computerlexikographie. In: HERBERT ERNST WIEGAND (Hg.): Wörterbücher in der Diskussion II. Tübingen 1996 (Lexicographica. Series Maior), 239–255.

STORRER 2001 = ANGELIKA STORRER: Digitale Wörterbücher als Hypertexte: Zur Nutzung des Hypertextkonzepts in der Lexikographie. In: INGRID LEMBERG/BERNHARD SCHRÖDER/ANGELIKA

STORRER (Hg.): Chancen und Perspektiven computergestützter Lexikographie. Tübingen 2001 (Lexicographica. Series Maior 107), 53–91.

WIEGAND 1997 = HERBERT ERNST WIEGAND: Über die gesellschaftliche Verantwortung der wissenschaftlichen Lexikographie. In: Hermes 18 (1997), 177–202.

WIEGAND 1998 = HERBERT ERNST WIEGAND: Wörterbuchforschung. Untersuchungen zur Wörterbuchbenutzung, zur Theorie, Geschichte, Kritik und Automatisierung der Lexikographie. 1. Teilbd.. Mit 159 Abbildungen im Text. Berlin, New York 1998.

WORTSCHATZ UNI LEIPZIG 2003 = Homepage WORTSCHATZ LEXIKON. <http://wortschatz.uni-leipzig.de/html/inhalt_geb.html>.

Alle Webseiten wurden am 1. Juni 2003 zum letzten Mal überprüft.