Sprachkorpora

Datenmengen und Erkenntnisfortschritt

Herausgegeben von
Werner Kallmeyer und Gisela Zifonun

Sonderdruck

Walter de Gruyter · Berlin · New York

Jörg Asmussen

Korpuslinguistische Verfahren zur Optimierung lexikalisch-semantischer Beschreibungen

Abstract

In diesem Beitrag wird untersucht, wie mithilfe korpuslinguistischer Verfahren Erkenntnisse über den Aufbau von Bedeutungsparaphrasen in Wörterbüchern gewonnen werden können. Diese Erkenntnisse sollen dazu genutzt werden, den Aufbau von Bedeutungsparaphrasen in Wörterbüchern umfassend und systematisch zu beschreiben, z. B. im Hinblick auf eine Optimierung der Bedeutungsparaphrasen für so genannte elektronische Wörterbücher oder für die Extraktion lexikalisch-semantischer Information für NLP-Zwecke.

1. Vorbemerkungen

Als Folge des Bestrebens, Sprachgebrauch lexikographisch immer präziser ermitteln und beschreiben zu können, ist die Verwendung von Korpora als Quelle bei der Erarbeitung insbesondere allgemeinsprachlicher Bedeutungswörterbücher der Gegenwartssprachen nicht mehr wegzudenken.

Unter *Korpus* ist in einem solchen Zusammenhang eine sehr umfangreiche, digitalisierte Sammlung von Texten oder Textauszügen zu verstehen, die als angenommene repräsentative Stichprobe eines durch verschiedene Parameter bestimmten Sprachgebrauchs dient, im allgemeinsprachlich-lexikographischen Kontext meist als intendierte Stichprobe der Allgemeinsprache schlechthin, vgl. zur Definition von *Korpus* in diesem Sinne einschlägige Beschreibungen und Diskussionen z. B. in Kennedy (1998), McEnery u. Wilson (2001), Tognini Bonelli (2001) oder Lemnitzer u. Zinsmeister (2006).

Wegbereiter für die Verwendung solcher Korpora in der lexikographischen Arbeit dürfte allen voran das englische *COBUILD*-Projekt gewesen sein, das 1980 als Kooperation zwischen der University of Birmingham und dem Verlag Collins begann, und dessen Ergebnis (COBUILD: Sinclair u. a. 1987) somit als das erste korpusgestützte Wörterbuch angesehen werden muss. Einige der korpuslinguistischen Verfahren, die diesem Wörterbuch zugrunde liegen, sind in Sinclair (1991) beschrieben.

Das erste und bisher einzige korpusgestützte Wörterbuch der dänischen Sprache ist *Den Danske Ordbog* (DDO: Hjorth u. a. 2005). Bei diesem Wörterbuch handelt es sich um ein sechsbändiges allgemeinsprachliches Bedeutungswörterbuch, vgl. Lorentzen (2004), das 1991–2005 von der Gesellschaft

für dänische Sprache und Literatur (*Det Danske Sprog- og Litteraturselskab*, DSL) von Grund auf neu erarbeitet und herausgegeben wurde.

Die Hauptquelle des DDO bildete das eigens zu diesem Zweck von der DSL zusammengestellte Textkorpus *Den Danske Ordbogs Korpus* (DDOC), vgl. Norling-Christensen u. Asmussen (1998). Es umfasst etwa 40 Millionen Textwörter in über 43000 Texten bzw. Textausschnitten einer Vielzahl von unterschiedlichen Textsorten aus dem Zeitraum 1983–1992; und es enthält detaillierte bibliographische sowie kommunikativ und soziolinguistisch orientierte Annotationen auf Textebene. Das DDOC diente bei der Erstellung der Stichwortliste für das DDO, der Ermittlung von Lesarten und ihrer Häufigkeiten, dem Auffinden von Kollokationen oder illustrativen Textbelegen sowie bei der Ermittlung orthographischer und morphologischer Varianz.

Die dem DDO zugrunde liegenden, durch Korpusevidenz weitgehend gesteuerten Verfahren dürften mittlerweile eine weite Verbreitung innerhalb der auf Gegenwartssprachen bezogenen (nicht ausschließlich kommerziell orientierten) lexikographischen Arbeit gefunden haben, so auch in deutschsprachigen Wörterbuchprojekten wie dem *elexiko* am Institut für Deutsche Sprache, vgl. Storjohann (2005), oder dem *DWDS* an der Berlin-Brandenburgischen Akademie der Wissenschaften, vgl. Klein (2004).

Bislang weniger verbreitet in der praktischen lexikographischen Arbeit ist die Verwendung von Korpora und korpuslinguistischen Verfahren bei der Untersuchung des Bedeutungsparaphrasenvokabulars oder des strukturellen Aufbaus von Bedeutungsparaphrasen. Dieser Beitrag wird sich daher nicht mit Einsatzmöglichkeiten von Korpora als *Quelle* für die lexikographische Sprachbeschreibung auseinander setzen, sondern wird vielmehr die Bedeutungsparaphrasen des Wörterbuchs selbst als ein spezielles Korpus betrachten, über das sich mithilfe allgemein üblicher korpuslinguistischer Verfahren Einblicke in den Aufbau von Bedeutungsparaphrasen gewinnen lassen. Das längerfristig angestrebte Ziel dabei ist es, eine umfassende systematische Beschreibung des Aufbaus von Bedeutungsparaphrasen zu erarbeiten. Auf einer solchen Grundlage ließen sich traditionelle, für das Buchmedium konzipierte Wörterbücher wie das DDO optimieren, indem sie den funktionalen Möglichkeiten des Mediums Computer besser angepasst werden könnten unter gleichzeitiger Berücksichtigung einer potenziellen Verwendbarkeit als NLP-Ressourcen.

2. Wörterbücher im digitalen Kontext

2.1 „State of the Art“

Im Laufe der Neunziger haben so genannte elektronische, also durch das Medium Computer vermittelte, Wörterbücher zunehmend an Verbreitung gewonnen. So sind mittlerweile von den meisten kommerziell produzierten Wörterbüchern auch digitale Versionen verfügbar. Das „Elektronische“ vieler dieser Wörterbücher besteht jedoch häufig lediglich darin, dass die her-

kömmliche Buchversion in eine inhaltlich und funktional völlig identische digitale Version umgesetzt wurde, die sich auf dem PC einsetzen lässt. Die Vorteile solcher elektronischer Wörterbücher beschränken sich daher in aller Regel auf Geschwindigkeitsgewinne beim Nachschlagen; qualitative Vorteile gegenüber der Buchversion, die außerdem das Funktionspotenzial des Mediums Computer wirklich ausschöpfen würden, scheint es bislang nur ansatzweise zu geben.

So verfügen einige englische Lerner-Wörterbücher in ihren jüngeren Versionen über Funktionalitäten, die scheinbar über diejenige herkömmlicher Buchversionen hinausreichen. Im *Longman Dictionary of Contemporary English* (LDOCE: Bullon u. a. 2003) gibt es die Funktion *Dictionary Search*, die es ganz ähnlich auch im *Macmillan English Dictionary* (MED: Rundell u. a. 2002) unter der Bezeichnung *SmartSearch* gibt. Beide erlauben es, scheinbar begriffsorientierte Recherchen durchzuführen, die bei näherer Untersuchung jedoch einige Defizite aufweisen. So ergibt beispielsweise eine Suche nach „yellow fruit“ im LDOCE – neben Wörtern, die wie zu erwarten eine gelbe Frucht denotieren – verblüffenderweise auch das Verb *bear*, während die Anfrage „instrument NOT string“ im MED neben solchen Wörtern, die tatsächlich saitenlose Instrumente bezeichnen, auch *banjo* zu Tage fördert. In Wirklichkeit wird keine semantische Suche vollzogen, sondern es wird lediglich untersucht, ob bei der Anfrage „yellow fruit“ die Zeichensequenzen *yellow* und *fruit* in ein und demselben Artikel vorkommen, während bei „instrument NOT string“ das Wort *instrument* in der Bedeutungsparaphrase vorkommen muss, das Wort *string* hingegen nicht vorkommen darf. Die Bedeutungsparaphrase des Lemmas *banjo* lautet *a musical instrument like a guitar but with a smaller round body*, weshalb es als vermeintlich saitenloses Instrument im Ergebnis der Suche erscheint. Bei näherer Betrachtung entpuppt sich die scheinbar begriffsorientierte Recherchefunktion als reine Volltextsuche, deren Skopus auf einen Artikel (LDOCE) bzw. bestenfalls auf die Bedeutungsparaphrasen (MED) beschränkt ist.

Trotz der Defizite zeigen die beiden Beispiele, dass ein Potenzial des Mediums Computer erkannt worden ist, nämlich die prinzipielle Möglichkeit einer onomasiologisch orientierten Recherche. Allerdings scheint das Datenmaterial der gedruckten Wörterbücher diesen neuen Möglichkeiten kaum (oder wahrscheinlich gar nicht) angepasst worden zu sein. Die Beispiele scheinen aber auch anzudeuten, dass eine mögliche Diskrepanz bestehen könnte zwischen dem Streben einer Wörterbuchredaktion nach verständlichen, allzu viel Redundanz und semantische Dekomposition vermeidenden Bedeutungserklärungen einerseits und der Notwendigkeit andererseits, entweder logisch exakte und stringente Bedeutungsparaphrasen in einem digitalen Kontext zu verwenden oder andere Verfahren zu entwickeln, die eine semantische Navigation in den herkömmlichen Wörterbuchdaten ermöglichen.

2.2 Ein Konzept für digitale Wörterbücher

Im Rahmen des DSL-Projektes *ordnet.dk* soll in einer ersten Phase 2004–2009 das ursprünglich nur für das Buchmedium konzipierte DDO in ein digitales lexikalisches Nachschlagewerk umgesetzt werden unter bewusster Berücksichtigung der Möglichkeiten für eine erweiterte Recherche, die das Computermedium prinzipiell bereitstellt. Insbesondere soll als eines der Ergebnisse der mit diesem Projekt verbundenen Entwicklungsarbeiten ein begriffsorientierter, onomasiologischer Zugriff auf den Datenbestand des Wörterbuchs ermöglicht werden.

Ausgangspunkt bei der Umsetzung ist die Annahme, dass ein „elektronisches Wörterbuch" im medialen Sinne kein Buch ist, sondern eine nach bestimmten funktionalen Kriterien organisierte lexikalische Datensammlung, weshalb man mit der Bezeichnung *digitales lexikalisches Nachschlagewerk* das Produkt vermutlich präziser beschreibt als mit der Bezeichnung *elektronisches Wörterbuch*.

Unterschiede zwischen einem Wörterbuch und einem digitalen lexikalischen Nachschlagewerk lassen sich in der Strukturierung des Datenmaterials und der Funktion, d. h. seiner intendierten und implizit gegebenen Verwendungsmöglichkeiten, feststellen.

So werden durch das Buchmedium vorgegebene strukturelle Bindungen aufgehoben, einmal hinsichtlich des Platzes, der prinzipiell nicht mehr beschränkt ist, zum anderen hinsichtlich der Artikelstruktur selbst, bei der eine Organisation des Materials nach funktionalen, oft sprachtechnologisch determinierten Kriterien für die Behandlung im Computer adäquater zu sein scheint, als eine unmittelbar auf den intendierten menschlichen Rezipienten ausgerichtete.

Weiter werden funktionale Bindungen des Buchmediums aufgehoben: Es muss nicht mehr vorab festgelegt werden, welche Funktionen ein geplantes Wörterbuch zu erfüllen hat. Dies eröffnet zumindest prinzipiell die Möglichkeit, die zugrunde liegenden lexikalischen Daten multifunktional anzulegen, z. B. auch im Hinblick auf eine eventuelle sprachtechnologische Verwendung. Erst bei der Präsentation wird das Datenmaterial in die jeweils entsprechende Form gebracht. Auch die typische funktionale Bindung eines ausschließlich alphabetisch indexierten, semasiologischen Zugriffs auf die Einträge wird prinzipiell aufgehoben.

2.3 Beispiel: Onomasiologischer Zugriff

Es soll somit also auf dem Hintergrund aufgehobener struktureller und funktionaler medialer Bindungen insbesondere der Frage nachgegangen werden, wie ein herkömmlicher lexikalischer Datenbestand mittels korpuslinguistischer Verfahren bearbeitet werden kann, sodass ein onomasiologischer Zugriff auf das Wörterbuch möglich wird. Dabei werden im Rahmen dieses Beitrags nur Substantive mit konkreten Lesarten berücksichtigt. Diese Beschränkung auf

die *1st order entities*, vgl. Lyons (1977), ist notwendig, um nicht den Rahmen dieses Beitrags zu sprengen, zumal die prinzipiell notwendige und interessante Erweiterung der Untersuchungen auf andere Wortarten und den *2nd* oder *3rd order entities* den Komplexitätsgrad der Beschreibung wesentlich erhöhen würde und den Fokus dieses Beitrags fort von den korpuslinguistischen Verfahren und hin zu der strukturellen Beschreibung von Bedeutungsparaphrasen verschieben würde.

Wenn ein semasiologisch angelegtes Wörterbuch um die Funktion eines onomasiologischen Zugriffs erweitert werden soll, bietet sich insbesondere die Analyse bestehender Bedeutungsparaphrasen an, um eine strukturelle Grundlage für einen solchen Zugriff zu erarbeiten. In diesem Zusammenhang sollte eine umfassende, generelle, systematische Beschreibung des strukturellen Aufbaus von Bedeutungsparaphrasen angestrebt werden; eine Beschreibung, die sich zumindest teilweise auf korpuslinguistische Verfahren stützen lässt, und die ein erster Schritt in Richtung auf eine eigentliche Formalisierung der lexikalisch-semantischen Information der Bedeutungsparaphrasen wäre als Voraussetzung für eine brauchbare algorithmische Prozessierung dieser Information.

In diesem Zusammenhang stellen sich bezüglich des DDO die folgenden beiden Fragen:

1. Welche korpuslinguistischen Verfahren ermöglichen einen Einblick in den Aufbau der Bedeutungsparaphrasen?
2. Welche Optimierungen der semantischen Beschreibung sind notwendig, um einen onomasiologischen Zugriff zu ermöglichen?

Vor der weiteren Erörterung dieser beiden Fragen in den Abschnitten 4 und 5 soll im Folgenden zunächst der Aufbau der semantischen Beschreibung im DDO näher beschrieben werden.

3. Semantische Beschreibung im DDO

3.1 Ein Beispiel

Im Folgenden wird der Aufbau der semantischen Beschreibung im DDO kurz dargelegt. Der Begriff *semantische Beschreibung* bezieht sich dabei auf denjenigen Teil der Mikrostruktur, in welchem insgesamt semantische Angaben zu einem Lexem gemacht werden.

Abbildung 1 zeigt einen Ausschnitt der semantischen Beschreibung des Lexems *mus* ‚Maus', nämlich das Semem *Nagetier*. Der hierarchische – in den Wörterbuchdaten durch XML ausgezeichnete – Aufbau der Mikrostruktur ist hier als etwas vereinfachte Baumstruktur wiedergegeben, wo die funktionalen lexikographischen Textelemente, vgl. Wiegand (1989a), die Blätter bilden, deren Funktion und Platzierung im strukturellen Gefüge aus den ihnen übergeordneten Knoten hervorgehen. Durch die Angabe eines Fragezeichens, Asteriskus oder Pluszeichens nach dem Knotennamen wird die mögliche Anzahl der von diesem Knoten ausgehenden Kanten angegeben,

wobei hier die für reguläre Ausdrücke üblichen Konventionen gelten. Fehlt die Angabe solcher Quantoren, kann von dem betreffenden Knoten genau eine Kante ausgehen. Sollten von einem Knoten null Kanten ausgehen, ist der Knoten obsolet. Solche Knoten, die also von der Strukturbeschreibung des Wörterbuchs her zwar mögliche, aber nicht realisierte Textelemente oder Teilstrukturen bezeichnen, sind in dieser Darstellung ausgelassen. In dem angeführten Beispiel sind bei weitem nicht alle von der formalen Mikrostrukturbeschreibung des Wörterbuchs her möglichen Angaben realisiert, sondern nur diejenigen, die für die semantische Beschreibung des Semems *Nagetier* im Artikel *mus* ‚Maus' relevant sind.

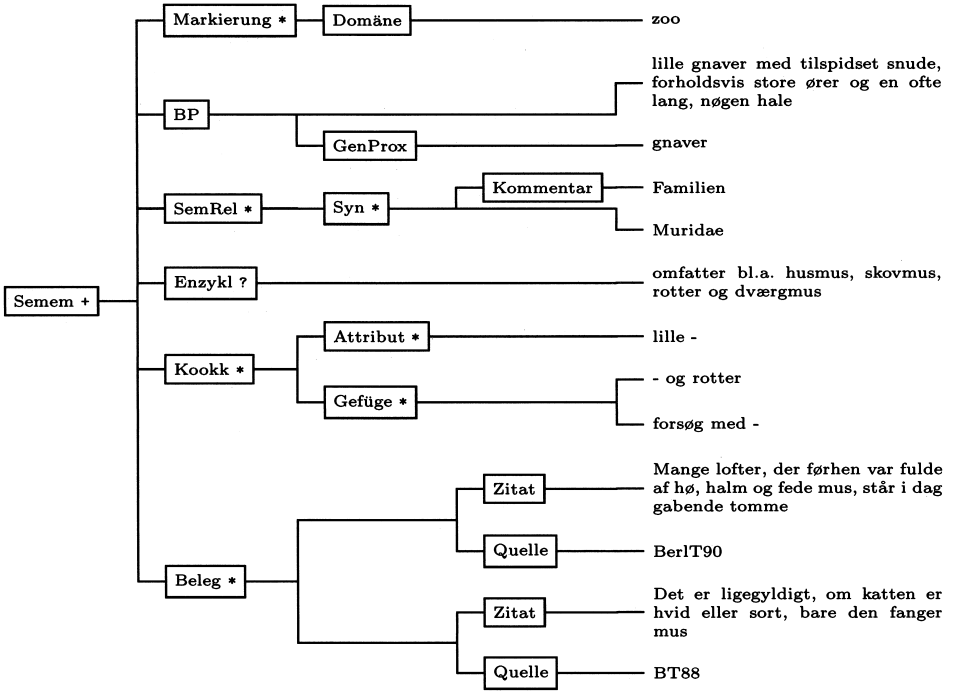

Abb. 1: Semantische Beschreibung im Artikel *mus* ‚Maus'

Unter dem Knoten *Markierung* gibt *Domäne* eine fachsystematische, für redaktionelle Zwecke genutzte und im Buch nicht mitgedruckte Domänenzuordnung des Semems an, hier *Zoologie*.

Die Angabe der Bedeutungsparaphrase *BP* besteht aus dem eigentlichen Textelement der im Wörterbuch gedruckten Bedeutungsparaphrase sowie unter dem Knoten *GenProx* aus der Angabe eines (nicht mitgedruckten) lexikalisierten Ausdrucks für das in der Bedeutungsparaphrase angegebene genus proximum.

Semantische Relationen wie Synonymie, Antonymie etc. werden unter dem Knoten *SemRel* gesammelt, im Beispiel eine durch die unter dem Knoten

Kommentar durch das Textelement *Familie* kommentierte Angabe eines Synonyms, nämlich der lateinischen Familienbezeichnung *Muridae*.

Unter dem Knoten *Enzykl* stehen enzyklopädische Angaben, in diesem Fall der Zusatz *omfatter bl.a. husmus, skovmus, rotter og dværgmus* ‚umfasst u. a. Hausmäuse, Waldmäuse, Ratten und Zwergmäuse', der sich auf die Synonymangabe zu beziehen scheint und somit übrigens einen Verstoß gegen das in diesem Wörterbuch verwendete XML-Auszeichnungsprinzip darstellt, wonach Daten von ihrer möglichen Repräsentation grundsätzlich zu trennen sind und somit der Inhalt einzelner Strukturelemente sich nicht durch eine präsupponierte typographische Abfolge auf diejenigen anderer beziehen sollte. Würde man z. B. beschließen, in einer modifizierten Version dieses Wörterbuchs – aus welchen Gründen auch immer – die Synonymangaben zu unterdrücken, ergäbe die hier im Strukturelement *Enzykl* gemachte Angabe nämlich keinen Sinn mehr.

Unter dem Knoten *Kookk* sind Angaben zu korpusstatistisch ermittelten Kookkurrenzen verzeichnet. Diese sind im gegebenen Beispiel wiederum unterteilt in attributive Kookkurrenzen unter dem Knoten *Attribut*, im Beispiel *lille (mus)* ‚kleine (Maus)' sowie *Gefüge* wie *(mus) og rotter* ‚(Mäuse) und Ratten' und *forsøg med (mus)* ‚Versuche an/mit (Mäusen)'.

Schließlich sind unter dem Knoten *Beleg* authentische, aus dem zugrunde liegenden Quellkorpus übernommene Verwendungsbeispiele angegeben, gegliedert in die Angabe des *Zitats* sowie seiner *Quelle*.

Von den relativ vielfältigen funktionalen Textelementen der semantischen Beschreibung im DDO sollen im Weiteren nur die eigentliche Bedeutungsparaphrase sowie die dazu gehörende Angabe des verwendeten genus proximum einer näheren Betrachtung unterzogen werden.

3.2 Die Bedeutungsparaphrase (BP)

3.2.1 Grundmuster: *genus proximum* und *differentia specifica*

Verglichen mit seinem Vorgänger, dem 28-bändigen, *Ordbog over det danske Sprog* (ODS: Dahlerup u. a. 1956), das mindestens sechs verschiedene Formen der Bedeutungsparaphrase (fortan auch BP) in einer Vielzahl struktureller Varianten verwendet, selbst in ein und demselben Artikel, vgl. Asmussen (2003), fallen die Bedeutungsparaphrasen im DDO wesentlich einheitlicher aus und gliedern sich in die Haupttypen: synonymische Paraphrasen, Verwendungsangaben sowie Paraphrasen nach dem klassischen Definitionsschema der Angabe des *genus proximum* sowie der *differentia specifica*. Die überwiegende Mehrheit der Bedeutungsparaphrasentypen – gegeben durch eine hohe Anzahl konkreter Substantive im Wörterbuch – ist nach diesem Muster aufgebaut.

Es sei an dieser Stelle darauf hingewiesen, dass das klassische Definitionsschema der Angabe eines genus proximum und der differentia specifica historisch dem Zweck der Bildung einer vorsprachlichen Ontologie diente und

nicht der Erklärung der Bedeutung sprachlicher Ausdrücke, vgl. Wiegand (1989b). Aufgrund seiner traditionellen Verankerung in der Lexikographie wird es im Folgenden zur Beschreibung eines Teilbereichs der lexikographischen Paraphrasierpraxis im DDO allerdings ganz bewusst verwendet, wobei ein genus proximum in diesem Kontext immer als ein Hyperonym zu verstehen ist, für den es (in der Sprache der Bedeutungsparaphrasen) einen *lexikalisierten Ausdruck* (in aller Regel in Gestalt eines Wortes) gibt: so meint die fortan verwendete Bezeichnung GP einen lexikalisierten Ausdruck für ein genus proximum.

Entsprechendes gilt für die differentia specifica (im Folgenden kurz DS), die hier zu verstehen ist als *sprachlicher Ausdruck* (bzw. eine Abfolge von sprachlichen Ausdrücken), der ein spezifizierendes Merkmal (bzw. ein Bündel solcher Merkmale) angibt.

Die Bedeutungsparaphrasen im DDO wurden nicht nach einem vorgegebenen semantischen Beschreibungsmodell abgefasst, in dem z. B. festgelegt gewesen wäre, welche semantischen Merkmale und Relationen bei welchen Wörtern z. B. in Abhängigkeit ihrer grundsätzlichen Semantik, z. B. den *entities* i. S. v. Lyons (1977), bzw. ihrer Wortart in der DS-Angabe beschrieben werden sollten. Auch wurde weder ein abgegrenzter Wortschatz für die Bedeutungsparaphrasen verwendet, noch gab es Vorgaben für ihren strukturellen Aufbau. Das Abfassen von Bedeutungsparaphrasen war somit dem einzelnen Redakteur relativ frei überlassen. Während es für die meisten anderen Elemente der Mikrostruktur explizit festgelegte Regeln gab, fehlten solche für das Abfassen von Bedeutungsparaphrasen fast völlig.

Auch wenn diese „pragmatisch“ orientierte Vorgehensweise beim Abfassen von Bedeutungsparaphrasen in der praktischen Lexikographie traditionell verankert ist und ihre Zweckmäßigkeit selten hinterfragt wird, so hat sie dennoch schwerwiegende Implikationen für die Nutzung solcher Bedeutungsparaphrasen in einem digitalen Kontext. Dies gilt insbesondere für die Identifikation der unterschiedlichen Elemente der differentia specifica, wo gleiche DS-Angaben sehr unterschiedlich zum Ausdruck kommen und auch sehr unterschiedlich in der Abfolge platziert sein können.

3.2.2 Das *genus proximum* (GP)

Die teilweisen Hyponymiestrukturen im DDO, die sich aus seinen in den Bedeutungsparaphrasen verwendeten genera proxima ermitteln lassen, besitzen auf der Grundlage der im vorangegangenen Abschnitt gegebenen Definition keine unbedingte Affinität zu einer vorsprachlichen Ontologie. Die Verwendung eines ausdrucksseitigen Kriteriums bei der Festlegung des GP ist hier motiviert durch die Notwendigkeit, GP-Angaben relativ zuverlässig algorithmisch bestimmen und verarbeiten zu können.[1]

[1] Auch die Einheiten, aus denen Korpora bestehen, sind zunächst nur ausdrucksseitig beschrieben (bzw. beschreibbar), wodurch sie algorithmisch prozessierbar werden.

Die Wahl eines geeigneten GP der BP lag bei der Ausarbeitung der DDO-Artikel stets im Ermessen des jeweils zuständigen Redakteurs, wobei man sich von einem nirgendwo näher explizitierten redaktionellen Ideal einer „allgemein verständlichen" BP leiten ließ. Die Verwendung einer vollends aus dem Quellkorpus eruierten Hyponymiestruktur wäre untersuchenswert, obwohl die allgemein hierfür verwendeten Methoden, vgl. z. B. Widdows (2003), eine Reihe von Unzulänglichkeiten aufweisen.

Zwei typische Bedeutungsparaphrasen aus dem DDO sind:

(1)
a. *grammofon* ‚Plattenspieler':

$[[_{GP}$ apparat $][_{DS}$ til at afspille grammofonplader med $]]_{BP}$
≈ Gerät zum Abspielen von Schallplatten

b. *cd-afspiller* ‚CD-Player':

$[[_{GP}$ apparat $][_{DS}$ til at afspille cd'er med $]]_{BP}$
≈ Gerät zum Abspielen von CDs

In beiden Bedeutungsparaphrasen ist als GP *apparat* ‚Apparat' bzw. ‚Gerät' angegeben, während die DS-Angabe sich – wie bei der Beschreibung von Artefakten sinnvoll – auf die Funktion bezieht, also die telische Rolle in der Qualiastruktur, vgl. Pustejovsky (1996), angibt. So auch im *Duden – Deutsches Universalwörterbuch* (DUW: Wermke u. a. 2003), das die folgenden Bedeutungsparaphrasen gibt:

(2)
a. *Plattenspieler*: $[[_{GP}$ Gerät $][_{DS}$ zum Abspielen von Schallplatten $]]_{BP}$

b. *CD-Player*: $[[_{GP}$ Abspielgerät $][_{DS}$ für Compact Discs $]]_{BP}$

Das eine Gerät dient also beiden Wörterbüchern zufolge dem Abspielen von Schallplatten, das andere dem Abspielen von CDs. Die Beispiele unter (1) und (2) zeigen, dass das DDO identische GP-Angaben in den Bedeutungserklärungen dieser beiden semantisch eng miteinander verwandten Begriffe hat, während das DUW im einen Fall das GP *Gerät*, im anderen das GP *Abspielgerät* verwendet. Zwar lässt sich dafür argumentieren, dass *Abspielgerät* lediglich eine lexikalisierte Variante der Phrase *Gerät zum Abspielen* ist, dass also in beiden Bedeutungsparaphrasen dasselbe GP in zwei sprachlichen Realisationen auftritt: das eine als *Wort* und das andere als *Phrase*. Da Phrasen laut hier verwendeter GP-Definition (vgl. Abschnitt 3.2.1) als GP allerdings nicht in Frage kommen, da sie sich algorithmisch nicht sicher genug ermitteln lassen, ist hier von zwei unterschiedlichen GP im DUW auszugehen, nämlich *Gerät* und *Abspielgerät*, wobei Letzteres ein Hyponym zu *Gerät* darstellen dürfte und deshalb selber das GP *Gerät* haben müsste. Im DUW fehlt allerdings der Eintrag *Abspielgerät*, sodass sich streng ausdrucksseitig-algorithmisch keine Hyperonyme hierzu im DUW ermitteln lassen.

Die identische GP-DS-Struktur in den Beispielen unter (1) deutet auf den ersten Blick darauf hin, dass das DDO ein höheres Maß an Konsistenz von Form und Inhalt der Bedeutungsparaphrasen erwarten lassen dürfte. Eine relative Konsistenz des Aufbaus der BP wurde im DDO dadurch erzielt, dass es weitgehend nach Bedeutungsfeldern redigiert wurde, d. h. dass sich typisch (aber nicht konsequent) nur *ein* Redakteur der Beschreibung der Ausdrücke eines semantischen Bereichs, z.B. aller Geräte, annahm und somit ein in sich einigermaßen schlüssiges Muster für die Bedeutungsparaphrasen konsequent verwenden konnte.

Trotzdem ist die auf den ersten Blick festzustellende Konsistenz keineswegs immer gewährleistet, wie folgende Beispiele zeigen:

(3)
a. *kanin* ‚Kaninchen':

[[$_{GP}$ gnaver][$_{DS}$ der ligner en lille hare, men har forholdsvis kortere ører og bagben og graver gange i jorden]]$_{BP}$

≈ Nagetier, das einem kleinen Hasen ähnelt, aber verhältnismäßig kürzere Ohren und Hinterbeine hat und Gänge in der Erde gräbt

b. *hare* ‚Hase':

[[$_{DS}$ gråbrunt][$_{GP}$ pattedyr][$_{DS}$ med hvid bug, to store fortænder i over- og undermunden, lange ører og kraftige bagben]]$_{BP}$

≈ graubraunes Säugetier mit weißem Bauch, zwei großen Schneidezähnen im Ober- und Unterkiefer, langen Ohren und kräftigen Hinterbeinen

c. *gnaver* ‚Nagetier':

[[$_{GP}$ pattedyr][$_{DS}$ med to store fortænder i over- og undermunden der er særligt egnede til at gnave med]]$_{BP}$

≈ Säugetier mit zwei großen Schneidezähnen im Ober- und Unterkiefer, die zum Nagen besonders geeignet sind

Im Beispiel (3-a) wird *kanin* ‚Kaninchen' mit dem GP *gnaver* ‚Nagetier' beschrieben, gleichzeitig wird die Ähnlichkeit mit *hare* ‚Hase' in der BP angegeben. *Hare* ‚Hase' selbst im Beispiel (3-b) erhält dahingegen das GP *pattedyr* ‚Säugetier' und in der DS-Angabe ein Merkmal, das auch Nagetiere auszeichnet (*store fortænder* ‚große Schneidezähne'). *Gnaver* ‚Nagetier' im Beispiel (3-c) ist schließlich ebenfalls als *pattedyr* ‚Säugetier' beschrieben. Genauso wie im Beispiel (2-b) ließe sich im Beispiel (3-b) semantisch-logisch durchaus für ein phrasal ausgedrücktes GP *pattedyr med to store fortænder i over- og undermunden* ‚Säugetier mit zwei großen Schneidezähnen im Ober- und Unterkiefer' als Äquivalent zum lexikalisierten *gnaver* ‚Nagetier' argumentieren. Es entspricht allerdings wiederum nicht der hier verwendeten Definition eines GP und muss von daher verworfen werden. In diesem Fall

würde sich die algorithmische Identifikation außerdem wesentlich schwieriger gestalten als im Beispiel (2-b), da die GP-Phrase in der BP unterbrochen wird durch eine DS-Angabe im Unterschied zur ganz entsprechenden GP-Phrase im Beispiel (3-c), die zusammenhängend (allerdings um einen Relativsatz erweitert) angegeben ist.

Diese Beispiele dürften zeigen, dass ein GP-Begriff, der außer lexikalisierten Einheiten auch Phrasen zuließe, sich in einem algorithmischen Kontext nur schwer operationalisieren lässt. Auf diesem Hintergrund stellt die Angabe des ferneren Hyperonyms *pattedyr* ‚Säugetier' (mit eventuell einschränkenden Merkmalen in den DS-Angaben) als GP für *hare* ‚Hase' unter gleichzeitiger Angabe des nächsten Hyperonyms *gnaver* ‚Nagetier' als GP für *kanin* ‚Kaninchen' eine Inkonsistenz in der Hyponymiestruktur dar. Solche Hyponymie-Inkonsistenzen sollten nach Möglichkeit im Wörterbuch bereinigt werden.

Abbildung 2 zeigt einen kleinen Ausschnitt aus der Hyponymiestruktur des DDO mit dem Ausgangspunkt im GP *gnaver* ‚Nagetier'.

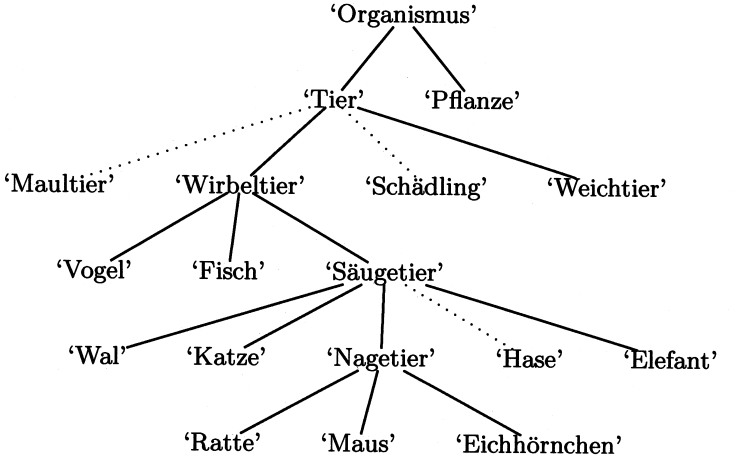

Abb. 2: Ausschnitt aus der impliziten Hyponymiestruktur des DDO

Abgesehen von der bereits besprochenen Hyponymie-Inkonsistenz geht aus der Abbildung weiter hervor, dass auch *muldyr* ‚Maultier' wahrscheinlich zu hoch in der Hierarchie angesiedelt wurde, obwohl andererseits das Hyperonym *dyr* ‚Tier' in der Allgemeinsprache in diesem Fall durchaus seine Berechtigung haben kann – und es ließe sich mit ziemlicher Wahrscheinlichkeit auf der Grundlage des Quellkorpus auch so belegen. Dieses Beispiel zeigt die Diskrepanz zwischen einer (fachsprachlichen) Taxonomie einerseits und (allgemeinsprachlichen) semantischen Beschreibungen im Wörterbuch andererseits – ein Umstand, dem bei einer algorithmischen Nutzung des Materials Rechnung getragen werden muss.

Aus dem abgebildeten Ausschnitt aus der Hyponymiestruktur im DDO geht weiter hervor, das in der BP von *skadedyr* ‚Schädling', das GP *dyr* ‚Tier' angegeben wurde. Nun ist ein Schädling zwar ein Tier und insofern ist das GP *dyr* durchaus gerechtfertigt, dennoch gehört es keiner bestimmten Art an und ist insofern auch kein Kohyponym zu *hvirveldyr* ‚Wirbeltier', *bløddyr* ‚Weichtier' oder gar *muldyr* ‚Maultier'; wäre dem so, würde dies unsinnigerweise ausschließen, dass ein Wirbeltier ein Schädling sein kann. Die Hyponymiebeziehung zwischen *skadedyr* ‚Schädling' und *dyr* ‚Tier' scheint demnach also eine qualitativ andere zu sein, als diejenige zwischen *hvirveldyr* ‚Wirbeltier' und *dyr* ‚Tier'. Dieser Unterschied sollte für eine algorithmische Nutzung des DDO explizitiert werden.

Weitere Schwierigkeiten für die algorithmische Nutzung des DDO-Materials ergeben sich aus dem Umstand, dass die GP-Angaben stets die Gestalt eines sprachlichen Ausdrucks haben, vgl. die Definition in Abschnitt 3.2.1, und nicht die eines (formal angegebenen) Inhalts, wie die Beispiele unter (4) und (5) belegen:

(4)
a. *svævefly* ‚Segelflugzeug':

[[$_{DS}$ motorløs][$_{GP}$ flyvemaskine][$_{DS}$ der skal trækkes op i luften (..) og derefter holder sig svævende (..)]]$_{BP}$
≈ motorloses Flugzeug, das (..) in die Luft hochgezogen wird und sich danach (..) in der Schwebe hält

b. *propelfly* ‚Propellerflugzeug':

[[$_{GP}$ fly][$_{DS}$ som drives frem vha. én el. flere propeller]]$_{BP}$
≈ Flugzeug, das durch einen oder mehrere Propeller vorangetrieben wird

(5)
a. *milt* ‚Milz':

[[$_{DS}$ aflangt mørkerødt][$_{GP}$ organ][$_{DS}$ i bughulen hvis vigtigste funktion er (..)]]$_{BP}$
≈ längliches, dunkelrotes Organ in der Bauchhöhle, dessen wichtigste Funktion es ist (..)

b. *magistrat* ‚Magistrat':

[[$_{GP}$ organ][$_{DS}$ som varetager den kommunale forvaltning i visse større byer (..)]]$_{BP}$
≈ Organ, das die kommunale Verwaltung in gewissen größeren Städten wahrnimmt (..)

In den beiden Beispielen unter (4) werden jeweils die synonymischen GP *flyvemaskine* und *fly* ‚Flugzeug' angegeben, was dazu führt, dass nach der rein ausdrucksseitigen GP-Auffassung hier zwei scheinbar unterschiedliche GP

verwendet werden, deren identischer Inhalt sich nicht unmittelbar algorithmisch ermitteln lässt. Interessanterweise wird dabei ausgerechnet bei *svævefly* ‚Segelflugzeug' die GP-Variante mit der Betonung der Motorisiertheit *flyvemaskine* wörtlich ‚Flugmaschine' angegeben.

In den Beispielen unter (5) wird das GP *organ* in zwei unterschiedlichen Lesarten verwendet, welches nach der rein ausdrucksseitigen GP-Auffassung dazu führt, dass *milt* ‚Milz' und *magistrat* in Übersichten wie in Abbildung 2 als Kohyponyme zu *organ* auftreten würden.

Die Beispiele zeigen somit, dass Synonymie und Polysemie bei GP-Angaben bei einer rein ausdrucksseitigen algorithmischen Nutzung des Wörterbuchmaterials zu Unregelmäßigkeiten führen werden, weshalb Synonymie und Polysemie bei GP-Angaben in einem digitalen Wörterbuch vermieden werden sollten.

Abschließend zeigt (6) ein Beispiel für ein GP mit nur einem Hyponym im ganzen Wörterbuch:

(6)
ildtang ‚Feuerzange':

[[$_{GP}$ jerntang][$_{DS}$ der bruges til at arrangere brænde i et ildsted]]$_{BP}$
≈ Eisenzange, die verwendet wird, um die Feuerung in einer Feuerstelle zurechtzurücken

Das verwendete GP *jerntang* ‚Eisenzange' hat nur das Hyponym *ildtang* ‚Feuerzange', ist also einmalig im gesamten DDO. Prinzipiell steht der algorithmischen Verwendung eines solchen Hapaxlegomenons nichts im Wege, solange man für das verwendete GP selbst wieder ein allgemeineres GP im Wörterbuch findet. In diesem Fall gibt es allerdings für *jerntang* ‚Eisenzange' ebenso wenig wie für *Abspielgerät* oben in Beispiel (2-a) einen eigenen Eintrag im Wörterbuch. *Ildtang* liegt somit außerhalb jeglicher Hyponymiestrukturen im DDO. Auch dieses müsste für eine algorithmische Nutzung des Materials behoben werden.

3.2.3 Die *differentia specifica* (DS)

Während die GP-Angaben durch ein gesondertes Element in der Artikelstruktur explizitiert werden, das zu Systematisierungszwecken, z. B. zur Bildung von teilweisen Begriffshierarchien wie in Abbildung 2, verwendet werden kann, lassen sich DS-Angaben aus den bestehenden Bedeutungsparaphrasen wesentlich schwieriger systematisieren.

Aus den Beispielen unter (1) geht hervor, dass die DS-Angaben sowohl von ihrer Semantik (in diesem Fall eine Funktionsbeschreibung) als von ihrem sprachlichen Ausdruck (*til at* ‚um zu' mit Infinitivphrase) her gleichermaßen aufgebaut sind. Dies legt die Vermutung nahe, dass sämtliche Bedeutungsparaphrasen hinsichtlich der DS-Angaben relativ einheitlich aufgebaut sind. Beispiel (7) zeigt allerdings, dass dem nicht so ist. Hier wird ebenfalls ein Lexem mit dem GP *apparat* beschrieben:

(7)
fjernsyn ‚Fernseher':

[[$_{\text{DS-1}}$ kasseformet][$_{\text{GP}}$ apparat][$_{\text{DS-2}}$ der kan modtage tv-signaler og omsætte dem til bevægelige billeder på en skærm og tilhørende lyd i apparatets højttalere]]$_{\text{BP}}$
≈ kastenförmiges Gerät, das TV-Signale empfangen und sie zu bewegten Bildern auf einem Schirm und dazugehörigem Ton in den Lautsprechern des Gerätes umsetzen kann

Unmittelbar fällt auf, dass in dieser BP auch Angaben zum allgemeinen Erscheinungsbild des Geräts gemacht werden. Der BP zufolge scheint die Kastenförmigkeit offenbar eine nennenswerte Eigenschaft eines Fernsehers zu sein, und offenbar scheint dieses Gerät im Gegensatz zu jenen im Beispiel (1) auch eine wesentlich detaillierte Funktionsbeschreibung verdient zu haben, was wenig plausibel erscheint. Darüber hinaus wird das GP in der DS in einer genitivischen Form wieder aufgenommen – was zur Komplexität dieser BP weiter beiträgt.

Es sind hier also zwei unterschiedliche Arten von differentiae specificae im Spiel: eine, die das Aussehen des Definiendums beschreibt, ausgedrückt als attributives Adjektiv, und eine weitere, die seine telische Rolle beschreibt und zwar in Gestalt einer finiten Form des Modalverbs *kunne* ‚können' mit zwei Infinitivphrasen *modtage tv-signaler* ‚TV-Signale empfangen' und *omsætte dem* ‚sie [TV-Signale] umsetzen'. Es kann auf dieser Grundlage die Annahme gemacht werden, dass viele Artefakt-Bedeutungsparaphrasen nach diesem Muster aufgebaut sind, also die fakultative Angabe einer Eigenschaft (DS-1) sowie die Angabe einer (oder mehrerer) Funktion(en) (DS-2) beinhalten. Ein nächster Schritt wäre, systematisch zu erkunden, wie diese DS-Angaben sprachlich in den Bedeutungsparaphrasen zum Ausdruck kommen können und auf dieser Grundlage eine systematische „Grammatik" der Bedeutungsparaphrasen zu formulieren. Diese könnte dann dazu eingesetzt werden, die Bedeutungsparaphrasen zu vereinheitlichen und dazu, quasi-algorithmisch bestimmte semantische Merkmale aus den Bedeutungsparaphrasen zu ermitteln, die z. B. für onomasiologische Recherchezwecke genutzt werden könnten.

Um jedoch einen Überblick über sämtliche Bedeutungsparaphrasen im DDO zu bekommen, bedarf es rationellerer Verfahren als des menschlichen Lesens, Analysierens und Beurteilens einer jeden BP im Wörterbuch. Hier können korpuslinguistische Verfahren hilfreich sein.

4. Korpuslinguistische Verfahren

Im Hinblick auf eine Anpassung des DDO-Datenbestandes an die Anforderungen eines onomasiologischen Zugriffs wurden in Abschnitt 2.3 die folgenden beiden Fragen aufgeworfen:

1. Welche korpuslinguistischen Verfahren ermöglichen einen Einblick in den Aufbau der Bedeutungsparaphrasen?
2. Welche Optimierungen der semantischen Beschreibung sind notwendig, um einen onomasiologischen Zugriff zu ermöglichen?

Auf dem Hintergrund der Untersuchungen im vorhergehenden Abschnitt soll im Folgenden einer möglichen Beantwortung dieser Fragen nachgegangen werden.

4.1 Wörterbuch und Korpus

Wie bereits in Abschnitt 1 angeführt wurde, so verwendet die herkömmliche korpusgestützte allgemeinsprachliche Lexikographie ein Korpus als angenommene repräsentative Stichprobe des im Wörterbuch zu beschreibenden Sprachgebrauchs. Eine Überlegung wäre nun, das Quellkorpus auch als Grundlage für den Ausdruck der lexikalisch-semantischen Beschreibung zu verwenden. Bedeutungsparaphrasen könnten so kontrollierbar allgemeinsprachlich abgefasst werden und nur mit tatsächlich im Korpus auch belegbaren Wörtern und sprachlichen Strukturen. So ließe sich mithilfe verschiedener korpuslinguistischer Verfahren das Ausdrucksinventar für die Formulierung von Bedeutungsparaphrasen festlegen.

Beispiele für diese Tendenz sind zum Teil *COBUILD* und *elexiko*. Zwar mag eine solche Annäherung der sprachreflexiven Sprache des Lexikographen an die Objektsprache, die sie semantisch beschreiben soll, dem menschlichen Wörterbuchbenutzer entgegenkommen, gleichzeitig dürfte sie aber, da sie nicht bewusst für ihren Zweck, nämlich Bedeutungen zu erklären, festgelegt wurde, eine algorithmische Prozessierung solchermaßen abgefasster Bedeutungsparaphrasen erschweren. Möchte man herkömmliche, nicht-formale Bedeutungsparaphrasen einem onomasiologischen Zugriff zugänglich machen, so sollte man im Gegenteil das sprachreflexive Ausdrucksrepertoire für Bedeutungsparaphrasen bewusst und eindeutig festlegen.

Die Hyponymierelationen, die den Bedeutungsparaphrasen zugrunde liegen sollen, ließen sich ebenfalls mit verschiedenen Mitteln aus Korpora ableiten. Ein solches Unterfangen wäre durchaus von einem gewissen Erkenntnisinteresse, da es einen Vergleich mit den implizit vorhandenen GP-Hierarchien im Wörterbuch erlauben würde und sicher zur Optimierung dieser Hierarchien beitragen könnte. Dennoch wären so aufgestellte Hierarchien wahrscheinlich von noch mehr Inkonsistenzen geprägt als die aus den Bedeutungsparaphrasen ableitbaren, weshalb auch hier ein bewusst festgelegtes sprachreflexives, lexikographisches Ausdrucksinventar zweckmäßiger erscheint als die Anlehnung an ein allgemeinsprachliches Korpus.

4.2 Wörterbuch als Korpus

Um mithilfe korpuslinguistischer Verfahren einen Einblick in den Aufbau von Bedeutungsparaphrasen (vgl. Frage 1) zu bekommen, werden diese nunmehr

als ein eigenständiges Spezialkorpus betrachtet. Das BP-Korpus des DDO ist dabei so strukturiert, dass jede BP im Korpus mit drei Attributen versehen ist, nämlich der Angabe des zu dieser BP gehörenden Schlagwortes, einer Angabe seiner Wortart sowie des in der BP verwendeten GP. So lassen sich z. B. gezielt BP mit bestimmten GP-Ausdrücken oder BP bestimmter Wortarten aus dem Korpus für Untersuchungszwecke herausfiltern. Das BP-Korpus des DDO umfasst knapp 87800 Bedeutungsparaphrasen mit insgesamt etwa einer Million Tokens.

Wie im Abschnitt 3 dargestellt wurde, ergeben sich bei näherer Untersuchung des Aufbaus von Bedeutungsparaphrasen einige Regelmäßigkeiten, die eine algorithmische Prozessierung nahe legen, aber auch eine ganze Reihe von Unregelmäßigkeiten, die eine algorithmische Prozessierung zumindest erschweren, wenn nicht gar ganz in Frage stellen.

Es soll daher nunmehr untersucht werden, mit welchen korpuslinguistischen Mitteln sich die unregelmäßigen Fälle auf der Grundlage des BP-Korpus generell für das gesamte Wörterbuch isolieren lassen, sodass eine anschließende Anpassung dieser Abweichungen an die mehrheitlich determinierte „Norm“ ermöglicht wird.

Wie die Untersuchungen im Abschnitt 3 gezeigt haben, fallen die Abweichungen in folgende Gruppen:

1. GP-Angaben:
   (a) Hyponymie-Inkonsistenzen, Beispiel *kanin* ‚Kaninchen‘ mit dem GP *gnaver* ‚Nagetier‘ vs. *hare* ‚Hase‘ mit dem GP *pattedyr* ‚Säugetier‘
   (b) Hyponymie-Diskrepanzen zwischen einer Taxonomie und Wörterbüchern, Beispiel *muldyr* ‚Maultier‘ mit dem GP *dyr* ‚Tier‘ und nicht *hvirveldyr* ‚Wirbeltier‘ oder *pattedyr* ‚Säugetier‘
   (c) Unterschiedliche Qualitäten von Hyponymiebeziehungen, Beispiel *skadedyr* ‚Schädling‘ vs. *hvirveldyr* ‚Wirbeltier‘ beide mit dem GP *dyr* ‚Tier‘
   (d) Synonymische GP-Ausdrücke, Beispiel *fly* vs. *flyvemaskine* beide ‚Flugzeug‘
   (e) Polyseme GP-Ausdrücke, Beispiel *organ*
   (f) GP mit nur einem Hyponym, Beispiel *jerntang* ‚Feuerzange‘
   (g) Wörterbuchintern hyperonymlose GP (die keine Wurzelknoten in der Hyponymiehierarchie bilden), Beispiel *jerntang* ‚Feuerzange‘

DS-Angaben:
   (a) Uneinheitliche Merkmalangaben
   (b) Uneinheitliche Verfahren in der syntaktischen Anordnung von Merkmalangaben

Mithilfe der nachfolgend aufgelisteten Reihe quantitativer korpuslinguistischer Verfahren ist es zum Teil möglich, einige dieser Unregelmäßigkeiten in den Bedeutungsparaphrasen zu erfassen.

1. Similaritätsprofile über Bedeutungsparaphrasen: Mithilfe verschiedener statistischer Similaritätsmaße ließe sich die oberflächliche Übereinstim-

mung von Bedeutungsparaphrasen z. B. hinsichtlich der Wahl und Anordnung der in ihnen verwendeten Tokens bestimmen. Hierbei ist anzunehmen, dass einander ähnliche Bedeutungsparaphrasen auch gemeinsame GP haben müssten. Werden nun einander ähnliche Bedeutungsparaphrasen mit unterschiedlichen GP gefunden, könnte dies womöglich auf eine Unregelmäßigkeit in der zugrunde liegenden semantischen Hierarchie hindeuten. Mithilfe einer solchen Methode ließen sich eventuell die unter (1-a) angeführten Unregelmäßigkeiten eingrenzen.

2. Frequenzprofile über das GP-Vokabular: Mithilfe solcher Frequenzprofile lassen sich vor allem niederfrequente GP ermitteln, die daraufhin durch solche auf höherem Niveau in der Hierarchie ausgetauscht werden können. Diese Methode eignet sich daher insbesondere zur Ermittlung der oben unter (1-f) angegebenen Unregelmäßigkeiten und zum Teil der unter (1-g) angeführten.
3. GP-Lemma-Abgleich: Bei diesem Verfahren werden zwei Frequenzprofile oder lediglich zwei Type-Inventarlisten erstellt: eine über GP und eine über die Lemmata des Wörterbuchs. Diese Listen werden anschließend miteinander verglichen, sodass GP, für die es im Wörterbuch kein Lemma gibt, herausgefiltert werden. Mithilfe dieses Verfahrens dürften sich insbesondere Unregelmäßigkeiten vom Typ (1-g) ermitteln lassen.
4. Frequenzprofile über das BP-Vokabular: Mithilfe dieses Verfahrens lassen sich insbesondere niederfrequente Wortformen ermitteln, die anschließend durch höherfrequente ersetzt werden könnten, was zu einer Vereinheitlichung des BP-Vokabulars (und der BP-Syntax) führen dürfte und somit zur Identifikation von Unregelmäßigkeiten der Typen (2-a) und (2-b) beitragen könnte.
5. Vergleiche von *n*-Gramm-Profilen: Weiter lassen sich signifikante syntagmatische Muster in den Bedeutungsparaphrasen durch Vergleiche von *n*-Gramm-Profilen des Bedeutungsparaphrasenkorpus mit solchen des Quellkorpus ermitteln. Als statistischer Test empfiehlt sich hierbei *Log Likelihood* (G^2), vgl. Dunning 1994. Auch dieses Verfahren kann Aufschlüsse über das Inventar und den Aufbau von DS-Angaben geben und lässt sich für Unregelmäßigkeiten der Typen (2-a) und (2-b) verwenden.
6. KWIC-Konkordanzen: Muster lassen sich anschließend mit einem Konkordanzwerkzeug für das gesamte Bedeutungsparaphrasenkorpus ermitteln, und es kann auf dieser Grundlage festgestellt werden, welche semantischen Spezifika durch welche Muster in den Bedeutungsparaphrasen zum Ausdruck gebracht werden. Auch dieses Verfahren dient insbesondere der Optimierung der DS-Angaben, Typen (2-a) und (2-b).
7. Ermittlung hervortretender DS-Types: Mithilfe eines statistischen Verfahrens zum Ermitteln unerwartet häufig auftretender Phänomene in einer bestimmten Situation (z. B. *Log Likelihood* oder *Mutual Information*, vgl. Church u. Hanks (1989)) lassen sich augenfällige DS-Types in Bedeutungsparaphrasen mit einem gemeinsamen GP ermitteln. Hierdurch lassen

sich Rückschlüsse auf bestimmte semantische Merkmale ziehen; außerdem lassen sich Unregelmäßigkeiten der Typen (2-a) und (2-b) so identifizieren.

Die hier aufgelisteten Methoden eignen sich überwiegend zur Optimierung der DS-Angaben in den Bedeutungsparaphrasen des Wörterbuchs. Unregelmäßigkeiten in der Hyponymiestruktur lassen sich hingegen wesentlich schwieriger mithilfe korpuslinguistischer Verfahren ermitteln. So beruhen die Methoden (1)–(3) nicht so sehr auf der Rohinformation im BP-Korpus, sondern vielmehr auf seiner GP-Annotation. Gäbe es diese nicht, wäre die Optimierung der Hyponymiestruktur noch schwieriger.

Im Folgenden soll nun beschrieben werden, wie einige der Methoden für die Optimierung von Bedeutungsparaphrasen im Wörterbuch eingesetzt werden können (vgl. Frage 2 oben).

5. Optimierung

Unter *Optimierung* ist hier die gezielte Anpassung der lexikalisch-semantischen Beschreibung insgesamt an bestimmte Anwendungsszenarien zu verstehen. Im Rahmen eines herkömmlichen Wörterbuchs schließt dies u. a. die bewusste, gezielte Formulierung von Bedeutungsparaphrasen im Hinblick auf den anvisierten Benutzer ein. Im Rahmen einer sprachtechnologischen Ressource wie beispielsweise eines WordNets sollte die Anpassung hingegen mögliche sprachtechnologische Anwendungen im Blick haben. In einem solchen Kontext wäre eine logisch konsistente Formalisierung der lexikalisch-semantischen Beschreibung sinnvoll. Hier ist das Anwendungsszenario zunächst einmal ein digitales Wörterbuch, das auch eine onomasiologische Recherche erlauben soll.

Ziel der Optimierung ist eine Verbesserung der Bedeutungsparaphrasen, worunter teils eine strukturelle Vereinheitlichung, teils überhaupt größtmögliche Konsistenz in der semantischen Beschreibung zu verstehen ist. Hierbei sollten sowohl unnötige Redundanz als auch unnötige Spezifizierungen weitgehend vermieden werden. Ziel ist es nicht, aus den bestehenden Bedeutungsbeschreibungen im DDO *unmittelbar* eine sprachtechnologische Ressource zu derivieren – derlei Unterfangen, ein maschinenlesbares herkömmliches Wörterbuch zu einer lexikalisch-semantischen NLP-Ressource zu konvertieren, wurden in den vergangenen etwa fünfundzwanzig Jahren unter Anwendung unterschiedlichster Methoden immer wieder unternommen und konnten bislang eher nur bescheidene Erfolge verzeichnen. Ide u. Véronis (1994) stellten bereits vor Jahren fest, dass die Derivation lexikalisch-semantischer Ressourcen aus Wörterbüchern erhebliche Hürden zu überwinden hat. Seither hat sich das Hauptaugenmerk der Forschung auf die Verwendbarkeit von *Korpora* für das teilweise automatisierte Erstellen sowohl herkömmlicher als auch sprachtechnologischer lexikalisch-semantischer Ressourcen gerichtet, so z. B. das von Kilgarriff u. Tugwell (2002) vorgestellte Verfahren, so genannte WordSketches aus Korpora abzuleiten.

Das Ziel der hier skizzierten Arbeit ist es also nicht, eine lexikalisch-semantische NLP-Ressource unmittelbar aus dem DDO abzuleiten, sondern zunächst einmal die lexikalisch-semantischen Beschreibungen, allen voran die Bedeutungsparaphrasen, im Wörterbuch selbst zu vereinheitlichen, um dann von hier aus eine Grundlage für eine mögliche zukünftige Generalisierung der semantischen Beschreibung zu schaffen, sodass letztendlich ein und derselbe lexikalische Datenbestand sowohl für gedruckte als auch für elektronische Wörterbücher genutzt werden kann und außerdem gleichzeitig sprachtechnologischen Anforderungen genügt.

5.1 GP-Angaben

Inkonsistenzen und Diskrepanzen in der Hyponymiestruktur lassen sich möglicherweise mithilfe des in Abschnitt 4.2 unter (2) dargestellten Verfahrens zum Teil identifizieren. Ein solches Verfahren ist augenblicklich bei der DSL in Entwicklung. Vorläufig erfolgt die Bereinigung von Inkonsistenzen und Diskrepanzen in der Hyponymiestruktur allerdings ohne Zuhilfenahme eines aus der Korpuslinguistik entlehnten quantitativen Verfahrens. Statt dessen werden Teilhierarchien aus dem Wörterbuch graphisch dargestellt und daraufhin von Redakteuren auf ihre Konsistenz hin untersucht und es werden offensichtlich notwendige Korrekturen vorgenommen.

Etwas einfacher stellt sich die Situation bei der Ermittlung und Bereinigung niederfrequenter GP dar. Von insgesamt 13748 unterschiedlichen GP-Types werden 6776 nur ein einziges Mal verwendet; Tabelle 1 zeigt einige solcher GP-Types, die überdies weder im zugrunde liegenden Quellkorpus vorkommen noch selbst Stichwörter im Wörterbuch bilden, die somit außerhalb der algorithmisch ermittelbaren Hyponymiestruktur des Wörterbuchs liegen.

GP-Type	Äquivalent
pergament*håndskrift*	‚Pergament*handschrift*‘
teater*figur*	‚Theater*figur*‘
stege*redskab*	‚Brat*werkzeug*‘
mund*del*	‚Mund*teil*‘
konkurrence*princip*	‚Wettbewerbs*prinzip*‘
minde*bygning*	‚Denkmal*gebäude*‘
farve*opløsning*	‚Farb*lösung*‘
kviksølv*forbindelse*	‚Quecksilber*verbindung*‘
symaskine*del*	‚Nähmaschinen*teil*‘
*lakrids*kugle	‚*Lakritz*kugel‘

Tab. 1: Einige GP-Types mit der Frequenz 1 im DDO

Bei den gezeigten GP-Types dreht es sich im Wesentlichen um okkasionell gebildete Komposita, die sich z. B. durch die kursivierten Teile in den Bedeutungsparaphrasen ersetzen ließen.

5.2 DS-Angaben

Da die DS-Angaben den größten Teil der Bedeutungsparaphrasen ausmachen, kommen korpuslinguistische Verfahren bei der Optimierung dieser Angaben besser zum Zug als bei den GP-Angaben. Im Folgenden werden nur die Möglichkeiten des in Abschnitt 4.2 unter (7) vorgestellten Verfahrens exemplarisch aufgezeigt.

Abbildung 3 zeigt die im DDO vorkommenden Lemmata, die in mindestens einer ihrer Bedeutungsparaphrasen das GP *gnaver* ‚Nagetier' haben. Untersucht man nunmehr sämtliche dieser Bedeutungsparaphrasen, so zeigt sich, dass bestimmte Merkmalsbezeichnungen in den DS-Angaben öfter wiederkehren – so z. B. ‚klein' bzw. ‚mausähnlich', ‚Haustier', ‚nicht einheimisch', ‚ohne Schwanz' – und zwar in den Bedeutungsparaphrasen derjenigen Lemmata, auf welche die vom entsprechenden Merkmal ausgehenden Pfeile in der Abbildung verweisen.

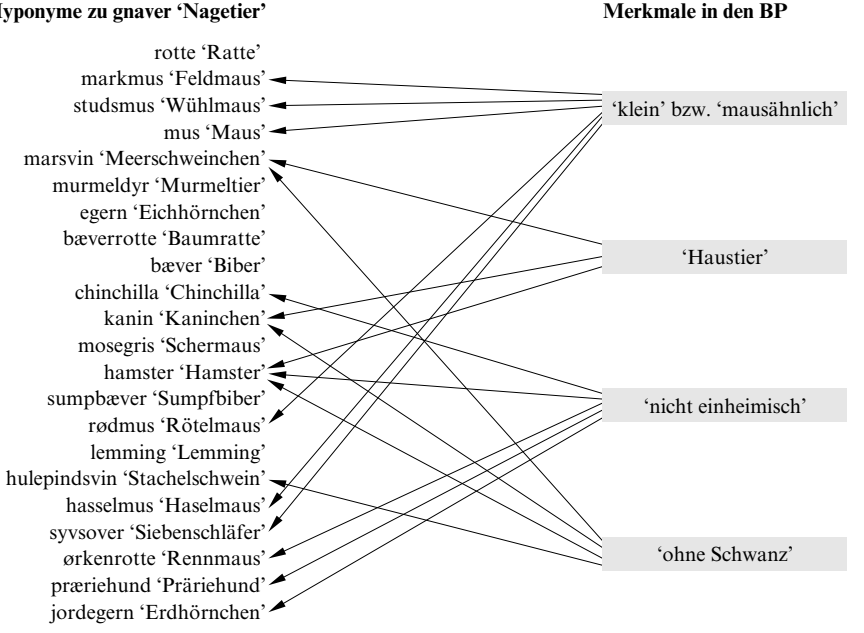

Abb. 3: Wiederkehrende Merkmale für *gnaver* ‚Nagetier'

Sämtliche Hyponyme zu *gnaver* ‚Nagetier' ließen sich auf der Grundlage derartiger Merkmalsuntersuchungen mit systematisch vereinheitlichten Angaben so ermittelter Merkmale versehen. Eine solche konsequent bearbeitete digitale Version des Wörterbuchs würde eine relativ zuverlässige Beantwortung einfacher onomasiologischer Anfragen ermöglichen wie *Welche Wörter bezeichnen Nagetiere, die als Haustiere gehalten werden?*

Bei Hyponymen zu *gnaver* ‚Nagetier', wovon es lediglich gut zwanzig im DDO gibt, ist eine manuell vorgenommene Gruppierung verschiedener Merkmale aus den DS-Angaben relativ überschaubar. Wesentlich unübersichtlicher gestaltet sich diese Aufgabe mit Hyponymen sehr hochfrequenter GP, wie den in Tabelle 2 gezeigten. Hier wäre für Gruppierungszwecke eine statistische Hilfe vonnöten.

GP-Type	Äquivalent	Häufigkeit
person	‚Person'	4382
del	‚Teil'	799
sted	‚Ort', ‚Stelle'	651
område	‚Gebiet'	422
gruppe	‚Gruppe'	421
plante	‚Pflanze'	389
måde	‚Art/Weise'	345
mængde	‚Menge'	326
stof	‚Stoff'	325
kvinde	‚Frau'	318
farve	‚Farbe'	314
genstand	‚Gegenstand'	291

Tab. 2: Häufige GP-Types im DDO

So können hervortretende DS-Types in Bedeutungsparaphrasen mit gemeinsamem GP z. B. durch die Verwendung des informationstheoretischen Maßes *Mutual Information* (MI) ermittelt werden, das sonst typisch für die Ermittlung von Kookkurrenzen auf der Grundlage eines bestimmten Ausgangswortes verwendet wird, vgl. Church u. Hanks (1989). In der hier vorgestellten Anwendung sind diese Ausgangswörter GP und die potenziell kookkurrierenden Wörter sämtliche Wortformen der Bedeutungsparaphrasen.

Im Einzelnen funktioniert die MI-basierte Ermittlung von DS-Types so, dass zunächst ein Korpus G aus sämtlichen Bedeutungsparaphrasen im Wörterbuch gebildet wird. Daraufhin wird dasjenige Teilkorpus T von G ermittelt, das aus Bedeutungsparaphrasen mit einem bestimmten gemeinsamen GP besteht.

Für jedes Type in einem solchen Teilkorpus T wird seine relative Frequenz im Teilkorpus f_T und seine relative Frequenz im Gesamtkorpus f_G ermittelt. Der Quotient aus diesen beiden Frequenzzahlen ergibt für jedes Type in T einen (gerundeten) Score $s \doteq \frac{f_T}{f_G}$. Je höher dieser Score, desto „enger" die Bindung zwischen dem jeweiligen Type und dem GP, das für T bestimmend ist. Der niedrigste Wert, den s annehmen kann, ist 1, der dann eintritt, wenn die Häufigkeiten des Types in T und G identisch sind. Dies wäre dann so zu interpretieren, dass es keine besondere Bindung zwischen diesem Type und dem untersuchten GP gibt.

Ein Beispiel: Für das Token *doven* ‚faul' in einem Korpus *T*, das durch das GP *person* gebildet wurde, ergibt sich der Score:

$$s = \frac{f_T\,(doven)}{f_G\,(doven)} = \frac{281}{19} \doteq 15$$

Vereinfacht ausgedrückt kommt *doven* etwa 15 mal häufiger in Bedeutungsparaphrasen mit dem GP *person* vor als in allen Bedeutungsparaphrasen des Wörterbuchs zusammengenommen.

Sortiert man eine so ermittelte Liste von Bedeutungsparaphrasentypes nicht-steigend nach ihren Scores, erhält man die für das jeweilige GP signifikantesten Bedeutungsparaphrasentypes an der Spitze dieser Liste.

Tabelle 3 zeigt die Spitze einer solchen Liste für das GP *gnaver* ‚Nagetier'; nämlich alle Types mit dem höchsten Score 2532. Der unmittelbare Eindruck von den aufgelisteten Types ist der, dass ihre Verbreitung in der Allgemeinsprache überwiegend wohl eher als gering einzuschätzen ist. In der Tat sind diese Types auch im BP-Korpus des DDO sehr niederfrequent: so kommen die Types *hårbeklædt* ‚haarbedeckt' und *nataktiv* ‚nachtaktiv' nur ein einziges Mal im Wörterbuch vor, nämlich in den Bedeutungsparaphrasen der Lemmata *studsmus* ‚Wühlmaus' und *hulepindsvin* ‚Stachelschwein'. Dass niederfrequenten Types ein so hohes Gewicht in der Berechnung des Scores beigemessen wird, ist eine grundlegende Eigenart von MI: Wenn ein bestimmtes Type z. B. nur wenige Male in *G* vorkommt und gleichzeitig wenige Male im betrachteten Teilkorpus *T*; wenn *T* überdies nur wenige Tokens enthält wie im Falle *gnaver* ‚Nagetier', dann wird der Score für ein sehr niederfrequentes, vielleicht nur zufällig im Wörterbuch verwendetes Token sehr hoch, ohne dass aus diesem Grund etwas für eine sonderlich „enge" Bindung zwischen dem Token und dem GP spräche. Diese Situation wird in Tabelle 3 widergespiegelt.

BP-Type	Äquivalent	Score
bæverlignende	‚biberähnlich'	2532
fodsåler	‚Fußsohlen'	2532
hårbeklædt	‚haarbedeckt'	2532
nataktiv	‚nachtaktiv'	2532
nutria	‚Nutria'	2532
prærieområder	‚Präriegebiete'	2532
Sibiriens	‚Sibiriens'	2532
tundra	‚Tundra'	2532
ungskov	‚Jungwald'	2532

Tab. 3: BP-Types mit höchstem Score für *gnaver* ‚Nagetier'

Unmittelbar scheint das MI-basierte Verfahren, hervortretende BP-Types für eine ganze Reihe von Kohyponymen zu ermitteln, seinen Zweck also zu verfehlen, denn niederfrequente Types bekommen scheinbar einen viel zu hohen

Score. Dies kann allerdings auch als Vorteil zur Optimierung von Bedeutungsparaphrasen genutzt werden. Denn es könnte durchaus etwas dafür sprechen, dass die so ermittelten Wortformen in den Bedeutungsparaphrasen überhaupt viel zu niederfrequent, zu speziell sind, als dass sie im Wörterbuch verwendet werden sollten.

So kommt keines der Types *bæverlignende* ‚biberähnlich', *hårbeklædt* ‚haarbedeckt', *prærieområder* ‚Präriegebiete' oder *ungskov* ‚Jungwald' im Quellkorpus DDOC vor, was als ein weiteres Indiz für ihre mangelnde Zweckmäßigkeit in den Bedeutungsparaphrasen gewertet werden könnte; und es ließe sich daher erwägen, diese durch geläufigere Formen zu ersetzen wie *som ligner en bæver* ‚der/die/das einem Biber ähnelt', *dækket af hår* ‚von Haaren bedeckt', *aktiv om natten* ‚aktiv nachts', *prærier* ‚Prärien' und *ung skov* ‚junger Wald'. Ein hierdurch eventuell erforderliches höheres Maß an syntaktischer Komplexität sowie hieraus resultierende Auswirkungen auf die Rezipierbarkeit der Bedeutungsparaphrasen müssten Gegenstand einer gesonderten Untersuchung sein.

Die niederfrequenten Types in den Ergebnissen des MI-Verfahrens sowie auch Formen mit einem zu niedrigen Score lassen sich mithilfe quantitativer Filter entfernen.[2] Niederfrequente Types lassen sich etwa mit der heuristisch ermittelten Frequenzschwelle $\log d+1$ eliminieren, wo d die Anzahl der Bedeutungsparaphrasen mit dem untersuchten GP ist. Ist die absolute Frequenz eines kookkurrierenden Types höher als der Schwellenwert, wird das Type bei der Ermittlung der hervortretenden Types weiter berücksichtigt, andernfalls nicht. Entsprechend lassen sich Types mit einem zu niedrigen Score durch die Schwelle $2 \cdot (\log d+1)$ eliminieren: hierbei werden nur Types berücksichtigt, deren Score über diesem Schwellenwert liegt. Tabelle 4 zeigt sämtliche so ermittelte Types in den Bedeutungsparaphrasen mit dem GP *gnaver* ‚Nagetier'. Ein Interpretationsansatz hierzu wäre, dass die durch die aufgelisteten

BP-Type	Äquivalent	Score
ører	‚Ohren'	342
hale	‚Schwanz'	194
pels	‚Fell'	140
gråbrun	‚graubraun'	127
korte	‚kurze'	55
lever	‚lebt'	53
lang	‚lang'	52
forholdsvis	‚verhältnismäßig'	24
kort	‚kurz'	14
lille	‚klein'	7

Tab. 4: Hervortretende BP-Types für *gnaver* ‚Nagetier' nach Filterung

[2] Die Verwendung eines statistischen Tests wie *Log Likelihood* anstatt MI wäre eine weitere Möglichkeit.

Types bezeichneten Merkmale zentral bei der Paraphrasierung von Bedeutungen mit dem GP *gnaver* ‚Nagetier' sind.

Es ließe sich nun für jede dieser Bedeutungsparaphrasen untersuchen, wie konsequent diese Merkmale erwähnt werden und in welchen syntaktischen Strukturen sie eingebettet sind: mangelnde Konsequenz ließe sich hierauf redaktionell bereinigen im Hinblick auf eine Vereinheitlichung des Aufbaus der Bedeutungsparaphrasen. So optimierte Bedeutungsparaphrasen ließen sich daraufhin besser algorithmisch behandeln. Eine solche Optimierung sei an einem weiteren Beispiel kurz aufgezeigt. Die Spitze der Liste der hervortretendsten Types in den Bedeutungsparaphrasen mit dem GP *kvinde* ‚Frau' ist in Tabelle 5 wiedergegeben.[3]

BP-Type	Äquivalent	Score
arrig	‚boshaft'	310
prostitueret	‚prostituiert'	241
stridbar	‚streitbar'	233
gift	‚verheiratet'	94
husholdning	‚Haushalt'	93
pige	‚Mädchen'	91
tiltrækkende	‚anziehend'	78
parforhold	‚Zweierbeziehung'	67
smuk	‚hübsch'	53
gaden	‚die Straße'	52

Tab. 5: Die zehn hervortretendsten BP-Types für *kvinde* ‚Frau'

Auf der Grundlage einer solchen Tabelle lassen sich Suchanfragen an das Bedeutungsparaphrasenkorpus formulieren, z. B. nach sämtlichen Bedeutungsparaphrasen, die das GP *kvinde* ‚Frau' und das DS-Type *prostitueret* ‚prostituiert' enthalten. Die hieraus resultierende KWIC-Konkordanz ist in Abbildung 4 dargestellt.

Da jede BP im Korpus mit der Angabe des zu ihr gehörigen BP-Ausdrucks und des Lemmas annotiert ist, gestaltet sich eine solche Anfrage ziemlich einfach.[4] In der wiedergegebenen Konkordanz ist die Annotation *Lemma* mit

[3] Die zehn hervortretendsten BP-Types des GP *mand* ‚Mann' sind übrigens *forfører* ‚verführt', *homoseksuel* ‚homosexuell', *uopdragen* ‚unerzogen', *kone* ‚Ehefrau', *adelig* ‚adelig', *uforskammet* ‚unverschämt', *parforhold* ‚Zweierbeziehung', *dreng* ‚Junge', *gift* ‚verheiratet' und *tiltrækkende* ‚anziehend'. Interessant wäre auf der Grundlage solcher quantitativ ermittelten Ergebnisse das in der Lexis verankerte Weltbild einmal näher zu untersuchen. In diesem Zusammenhang wäre ein Vergleich des Weltbildes in den redaktionell erstellten Bedeutungsparaphrasen mit einer entsprechenden quantitativen Untersuchung der dem Wörterbuch zugrunde liegenden Korpusdaten interessant im Hinblick auf die Frage, ob Wörterbuchredakteure das im Korpus vermittelte sprachliche Weltbild unmittelbar oder doch eher introspektiv in den Bedeutungsparaphrasen des Wörterbuchs vermitteln.

[4] Als Anfragesystem wird die ursprünglich am Institut für Maschinelle Sprachverarbei-

angezeigt. Es fällt unmittelbar auf, dass eine Reihe von Bedeutungsparaphrasen nach folgendem Muster aufgebaut sind:

(8)
$[[_{DS}$ prostitueret $][_{GP}$ kvinde $][_{DS} \ldots]]_{BP}$
≈ prostituierte Frau …

[genus="kvinde" & norm="prostitueret"]

lemma	left context	match	right context
kurtisane	Elegant	prostitueret	kvinde med dannelse god smag og g
hore	Kvinde som bedriver hor	prostitueret	kvinde
demimonde	Kvinde der helt el delvist ernærede sig som	prostitueret	
massøse		Prostitueret	kvinde på en massageklinik
hetære		Prostitueret	kvinde i antikkens Grækenland ofte r
glædespige		Prostitueret	kvinde
gadetøs		Prostitueret	kvinde der trækker på gaden
gadepige		Prostitueret	kvinde der trækker på gaden
fruentimmer		Prostitueret	kvinde
callgirl		Prostitueret	kvinde der tilkaldes telefonisk ofte sc
bajadere		Prostitueret	kvinde

Abb. 4: BP-Konkordanz

Insgesamt gibt es im Wörterbuch elf Lemmata mit dieser Lesart, bei denen das unter (8) gezeigte BP-Muster verwendet wird.

Bei näherer Untersuchung der zu den weggefilterten DS-Types gehörenden Lemmata zeigt sich, dass unter ihnen weitere mit der Lesart ‚prostituierte Frau' auftreten z. B. *ludder* und *skøge* – beide überdies sogar monosem. Die zu diesen beiden Beispielen gehörenden Bedeutungsparaphrasen sind unter (9) wiedergegeben: deutlich zeigt sich, dass hier zwei Abweichungen vom Muster unter (8) vorliegen.

(9)
a. *ludder* ‚Nutte':

$[[_{GP}$ kvinde $][_{DS}$ som tilbyder samleje og andre seksuelle ydelser mod betaling $]]_{BP}$
≈ Frau, die Geschlechtsverkehr und andere sexuelle Leistungen gegen Bezahlung anbietet

b. *skøge* ‚Dirne':

$[[_{GP}$ kvinde $][_{DS}$ der ernærer sig ved prostitution $]]_{BP}$
≈ Frau, die sich durch Prostitution ernährt

Somit eignet sich dieses quantitative Verfahren zum Aufdecken von Unregelmäßigkeiten in der Struktur von an sich verwandten Bedeutungsparaphrasen.

---

tung der Universität Stuttgart entwickelte IMS Corpus Workbench (CWB) verwendet, vgl. Christ 1994. Eine Open-Source-Version hiervon wird unter http://cwb.sourceforge.net/ zugänglich gemacht.

Eine auf dieser Grundlage systematisch durchgeführte Optimierung der Bedeutungsparaphrasen – möglicherweise unter Zuhilfenahme weiterer korpuslinguistischer Verfahren ließe sich somit für eine Vereinheitlichung der Bedeutungsparaphrasenstruktur im Wörterbuch verwenden. Solchermaßen optimierte oder standardisierte Bedeutungsparaphrasen erleichtern möglicherweise die Rezeption durch den menschlichen Benutzer, vor allem ermöglichen sie aber eine zuverlässigere algorithmische Auswertung als die nicht optimierten Bedeutungsparaphrasen, die in aller Regel in herkömmlichen Wörterbüchern anzutreffen sind. Von daher wäre ein so bearbeitetes Wörterbuch in einem digitalen Kontext wahrscheinlich funktionstüchtiger als die im Abschnitt 2.1 vorgestellten digitalen Versionen gedruckter Wörterbücher.

6. Ausblick

Optimierungen der Bedeutungsparaphrasen, wie sie hier ansatzweise beschrieben wurden, würden zwar das Wörterbuch als Grundlage onomasiologischer Anfragen zuverlässiger machen, dennoch würden weiterhin Unwägbarkeiten der freien Formulierung der Bedeutungsparaphrasen fortbestehen, die eine algorithmische Nutzung des Materials weiterhin relativ schwierig gestalten würden. Es scheint somit ein nur schwer zu überwindender Widerspruch zu bestehen zwischen den Bedürfnissen menschlicher Wörterbuchbenutzer, die rezipierbare Bedeutungsparaphrasen erwarten, und algorithmischen Systemen, die eine schlüssige, logische Struktur der semantischen Beschreibung im Wörterbuch erfordern. Diese Diskrepanz ließe sich zum Teil überwinden, wenn auf der Grundlage korpuslinguistischer Analysen der Bedeutungsparaphrasen zunächst einmal eine eher formal strukturierte lexikalisch-semantische Ressource entwickelt würde, die sich mit den semantischen Beschreibungen im bestehenden Wörterbuch verknüpfen ließe. Dieses Ziel verfolgt das zunächst bis Ende 2008 laufende Projekt *DanNet*, vgl. Pedersen u. a. (2006), das als Kooperation zwischen der DSL und dem *Center for Sprogteknologi*, *CST*, ein WordNet für das Dänische entwickelt, vergleichbar etwa dem Princeton WordNet, vgl. Fellbaum (1998), oder den verschiedenen EuroWordNets, vgl. Vossen (1999), jedoch ergänzt durch Angaben zur Qualiastruktur, vgl. Pustejovsky (1996). Die WordNet-Entwicklung erfolgt auf der Grundlage des DDO und unter Einsatz der in diesem Beitrag geschilderten Verfahren. Da dieses WordNet unmittelbar vom DDO abgeleitet wird, kann es als eine algorithmisch verwertbare Parallelversion zum Wörterbuch betrachtet werden, und ist als solche auch unmittelbar mit ihm verknüpft, indem die aus ihm abgeleiteten Synsets unmittelbar mit entsprechenden Bedeutungsparaphrasen im DDO verbunden sind. Auf dieser Grundlage dürften die angestrebten onomasiologisch orientierten Recherchen in der künftigen digitalen Version des DDO von einer besseren Qualität sein als solche, die sich ausschließlich auf eine optimierte semantische Beschreibung im Wörterbuch stützen würden.

Ein weiterer Vorteil dieses Verfahrens ist, dass zunächst nicht in die bestehende semantische Beschreibung im DDO eingegriffen wird. Gleichzeitig ermöglicht es eine gründliche Analyse und formale Umsetzung der in den Bedeutungsparaphrasen enthaltenen Informationen.

So wird die Bereinigung von Inkonsistenzen und Diskrepanzen in der Hyponymiestruktur zunächst einmal nur im Rahmen des WordNets durchgeführt. Erst wenn hier eine akzeptable Hyponymiestruktur erreicht worden ist, wird auf dieser Grundlage dann in einem weiteren Schritt die Anpassung der GP in den Bedeutungsparaphrasen des Wörterbuchs durchgeführt. In dem Prozess der Bereinigung der Hyponymiestruktur werden die GP, die im DDO sowie in diesem Beitrag als GP bezeichnende Ausdrücke verstanden sind (vgl. Abschnitt 3.2.1) durch die von ihnen bezeichneten *Inhalte* ersetzt, was das Problem synonymer und polysemer GP-Ausdrücke löst. Auch qualitativ unterschiedliche Hyperonymierelationen werden entsprechend markiert, vgl. hierzu Pedersen u. Sørensen (2006).

Auf der Grundlage der DS-Analysen, z. B. der quantitativen Ermittlung hervortretender DS-Angaben, wird nicht nur eine Umsetzung der semantischen Information für DanNet vorgenommen, sondern als weiteres Ziel auch eine systematische Beschreibung der Mikrostruktur von Bedeutungsparaphrasen überhaupt angestrebt. Somit ließen sich unterschiedlich aufgebaute Bedeutungsparaphrasen bereits bei der Ausarbeitung eines Wörterbuchs weitgehend vermeiden. Eine solche systematische Strukturbeschreibung von Bedeutungsparaphrasen könnte in Kombination mit einer WordNet-Struktur wiederum die Grundlage für eine formale Beschreibung der lexikalischen Semantik hinter den eigentlichen Bedeutungsparaphrasen bilden – eine Beschreibung, die letztendlich anstelle der klassischen Bedeutungsparaphrasen treten könnte: diese würden dann je nach Bedarf aus den formalen Beschreibungen abgeleitet werden. Somit ließe sich auch die sprachreflexive Sprache der Bedeutungsparaphrasen zunächst einmal vermeiden. An ihrer Stelle würde dann intern eine eigentliche semantische Metasprache treten, die auf den jeweiligen Verwendungszweck des Wörterbuchs ausgerichtet in eine adäquate sprachreflexive Sprache umgesetzt werden könnte. So ließe sich auch das Problem, dass der sprachliche Aufbau und Ausdruck in den Bedeutungsparaphrasen eines Wörterbuchs möglicherweise nur ein zeitlich begrenztes Modephänomen darstellt, leichter umgehen.

Die in diesem Beitrag lediglich angerissenen Themen dürften einen Eindruck davon vermittelt haben, dass die Verwendung von Korpora und korpuslinguistischen Verfahren in der Lexikographie nicht nur einen Erkenntnisfortschritt für denjenigen Teil der lexikographischen Arbeit bedeuten, wo ein Korpus als Quelle für die lexikographische Beschreibung verwendet wird, sondern auch dort, wo sie als Werkzeug bei der lexikographischen Arbeit mit dem Wörterbuch selbst zum Einsatz kommen.

Ein digitales Wörterbuch, das wirklich die Möglichkeiten des Mediums Computer ausschöpfen will, sollte in seiner Konzeption eine multifunktionale

lexikalische Ressource sein, die sowohl in unterschiedlichen Buchversionen erscheinen kann als auch als digitales Nachschlagewerk mit unterschiedlichen Recherchemöglichkeiten – und die sich überdies auch für NLP-Zwecke verwenden ließe.

Literatur

Asmussen, Jörg (2003): Zur geplanten Retrodigitalisierung des Ordbog over det danske Sprog. Konzeption, Vorgehensweise, Perspektiven. In: Burch, Fournier, Gärtner u. Rapp (Hg.) 2003: Standards und Methoden der Volltextdigitalisierung. Beiträge des Internationalen Kolloquiums an der Universität Trier, 8./9. Oktober 2000 (Abhandlungen der Akademie der Wissenschaften und der Literatur). Mainz.

Christ, Oliver (1994): A modular and flexible architecture for an integrated corpus query system. In: Proceedings of COMPLEX'94. Budapest.

Church, Kenneth u. Hanks, Patrick (1989): Word association norms, mutual information and lexicography. In: ACL Proceedings, 27th Annual Meeting. Vancouver.

COBUILD: Sinclair, John u. a. (Hg.) (1987): Collins COBUILD English Language Dictionary. London.

DDO: Hjorth, Kristensen, Lorentzen, Trap-Jensen, Asmussen u. a. (Hg.) (2005): Den Danske Ordbog 1–6. København.

Dunning, Ted (1994): Accurate Methods for the Statistics of Surprise and Coincidence. In: Computational Linguistics 19(1), S. 61–74.

DUW: Wermke, Matthias u. a. (Hg.) (2003): Duden Deutsches Universalwörterbuch. 5. überarbeitete Auflage. Mannheim.

Fellbaum, Christiane (Hg.) (1998): WordNet: An Electronic Lexical Database. Cambridge, MA.

Ide, N. u. Véronis, J. (1994): Knowledge Extraction from Machine-Readable Dictionaries: An Evaluation. In: Steffens, P. (Hg.): Machine Translation and the Lexicon. Springer-Verlag.

Kennedy, Graeme (1998): An Introduction to Corpus Linguistics. London.

Kilgarriff, Adam u. Tugwell, David (2002): Sketching Words. In: Corréard, M.-H. (Hg.): Lexicography and Natural Language Processing. A Festschrift in Honour of B. T. S. Atkins. Euralex.

Klein, Wolfgang (2004): Das digitale Wörterbuch der deutschen Sprache des 20. Jahrhunderts. In: Scharnhorst J.: Sprachkultur und Lexikographie. Berlin.

LDOCE: Bullon, Stephen u. a. (Hg.) (2003): *Longman Dictionary of Contemporary English*. Pearson Education.

Lemnitzer, Lothar u. Zinsmeister, Heike (2006): Korpuslinguistik. Tübingen.

Lorentzen, Henrik (2004): The Danish Dictionary at large: presentation, problems and perspectives. In: Proceedings of the 11th EURALEX International Congress. Lorient.

Lyons, John (1977): Semantics. Cambridge.

McEnery, Tony u. Wilson, Andrew (2001): Corpus Linguistics. Edinburgh.

MED: Rundell, Michael u. a. (Hg.) (2002): Macmillan English Dictionary for Advanced Learners. Oxford.

Norling-Christensen, Ole u. Asmussen, Jørg (1998): The Corpus of The Danish Dictionary. In: Lexikos. Afrilex Series 8.

ODS: Dahlerup, Verner u. a. (Hg.) (1956): Ordbog over det danske Sprog 1–28. København.

Pedersen, Bolette S., Nimb, Sanni, Asmussen, Jørg u. a. (2006): DanNet – a WordNet for Danish. In: Proceedings of the Third International WordNet Conference. Jeju, Korea.

Pedersen, Bolette S. u. Sørensen, Nicolai H. (2006): Towards Sounder Taxonomies in WordNets. In: Proceedings from the OntoLex Workshop in association with LREC 2006. Genova.

Pustejovsky, James (1996): The Generative Lexicon. Reprint. Bradford Book.

Sinclair, John (1991): Corpus, Concordance, Collocation. Oxford.

Storjohann, Petra (2005): elexiko: A Corpus-Based Monolingual German Dictionary. In: Hermes, Journal of Linguistics 34. Aarhus.

Tognini Bonelli, Elena (2001): Corpus Linguistics at Work. Amsterdam/Philadelphia.

Vossen, Piek (Hg.) (1999): EuroWordNet. A Multilingual Database with Lexical Semantic Networks. Amsterdam.

Widdows, Dominic (2003): Geometry and Meaning. Center for the Study of Language and Information – Lecture Notes (CSLI-LN). Chicago.

Wiegand, H. E. (1989a): Der Begriff der Mikrostruktur: Geschichte, Probleme, Perspektiven. In: Hausmann, Reichmann, Wiegand, Zgusta: Wörterbücher. Ein internationales Handbuch zur Lexikographie. Berlin/New York.

Wiegand, H. E. (1989b): Die lexikographische Definition im allgemeinen einsprachigen Wörterbuch. In: Hausmann, Reichmann, Wiegand, Zgusta: Wörterbücher. Ein internationales Handbuch zur Lexikographie. Berlin/New York.