Simulation lexikalischen Erwerbs

James Kilbury, Petra Naerger, Ingrid Renz*

Zusammenfassung

Im folgenden wird ein Ansatz vorgestellt, der die Verarbeitung unbekannter Wörter und den Erwerb entsprechender Lexikoneinträge modelliert. Diese Modellierung, die durch linguistisch motivierte Grundannahmen bestimmt ist, wurde im Rahmen des unifikationsbasierten Paradigmas als Computersimulation in Form des Systems QPATR realisiert. Ausgehend von der zentralen Annahme, daß die Verarbeitung unbekannter Wörter denselben Prinzipien unterliegt wie die Verarbeitung natürlicher Sprache im allgemeinen, wird gezeigt, wie Information über unbekannte Wörter bei der Analyse gewonnen werden kann. Während dieser Aspekt des lexikalischen Erwerbs somit keine zusätzlichen Mittel erfordert, ist für die Erstellung von neuen Lexikoneinträgen eine eigenständige Komponente notwendig. Diese Komponente, die auf einem expliziten Lexikonmodell basiert, bildet nicht nur neue Lexikoneinträge, sondern integriert diese auch in das bestehende Lexikon.

1 Generelle Zielsetzung

Das Ziel unserer Arbeit ist es, in einem implementierten System den Erwerb lexikalischer Einheiten zu modellieren. Auf diese Weise werden spezifische Leistungen der menschlichen Sprachfähigkeit im Rahmen der Computerlinguistik explizit modelliert.

Für die maschinelle Verarbeitung von Sprache ergibt sich als Konsequenz, daß eine zu analysierende Eingabe Wörter enthalten kann, für die es im Lexikon keine geeigneten Lexikoneinträge gibt und die somit dem sprachverarbeitendem System unbekannt sind. Zunächst muß das System fähig sein, zu erkennen, daß ein Wort einer Eingabekette keinen geeigneten Lexikoneintrag besitzt. Auch in diesem Fall muß das

* Die hier vorgestellte Arbeit wurde im Projekt SIMLEX ("Simulation lexikalischen Erwerbs") entwickelt. Das Projekt wird von der Deutschen Forschungsgemeinschaft gefördert (vom 1.7.89 - 31.12.91 im Schwerpunkt "Kognitive Linguistik", seit 1.1.92 im Sonderforschungsbereich 282 "Theorie des Lexikons"). Für anregende Diskussionen und hilfreiche Kommentare danken wir Gosse Bouma, Lynne Cahill, Roger Evans, Gerald Gazdar und Dafydd Gibbon.

System eine angemessene Verarbeitung der Eingabekette leisten. Darüberhinaus müssen auch die in der Eingabe enthaltenen unbekannten Wörter aufgrund ihres sprachlichen Kontextes analysiert werden. So könnte beispielsweise für das angenommene unbekannte Wort *Knuffel* in dem Satz *der Knuffel schläft* aus dem Kontext geschlossen werden, daß es sich um ein Substantiv handelt mit den syntaktischen Merkmalen **dritte Person singular maskulin nominativ** und dem semantischen Merkmal **belebt**. Schließlich müssen für die unbekannten Wörter aufgrund der ihnen zugeordneten Analysen neue Lexikoneinträge erstellt werden und diese in das bestehende Lexikon integriert werden.[1]

Verschiedene theoretische Grundannahmen schränken die möglichen Lösungsansätze ein. Die Hauptannahme ist dabei, daß die Analyse unbekannter Wörter denselben Prinzipien unterliegt wie die Verarbeitung natürlicher Sprache im allgemeinen. Dieses Postulat verhindert die Einführung neuer Systemkomponenten für die Analyse unbekannter Wörter, die speziell für diese Aufgabe konzipiert sind. Es erlaubt hingegen die Modifikation bestehender Komponenten. Außerdem betrifft diese Annahme nur die Analyse, nicht aber die Erstellung der neuen Lexikoneinträge. Da diese Arbeit nur bei unbekannten Wörtern geleistet werden muß, ist dafür eine eigenständige Systemkomponente zu entwickeln.

Eine Annahme, die den Forschungsgegenstand eingrenzt, ist, daß die Simulation des lexikalischen Erwerbs auf einen idealisierten erwachsenen Leser beschränkt ist. Diese Annahme grenzt Fragestellungen des Erstspracherwerbs aus. Während ein Kind lexikalisches und grammatisches Wissen zugleich erfassen muß, verfügt ein idealisierter erwachsener Leser bereits über vollständiges grammatisches Wissen, das sich auch durch eine Erweiterung des lexikalischen Wissens nicht verändert. Dagegen ist das Lexikon dieses idealisierten Lesers charakteristischerweise unvollständig und wird ständig erweitert.

Gemäß dem zu modellierenden Lexikon des idealisierten Lesers wird auch das Lexikon des Systems als prinzipiell unvollständig und erweiterbar betrachtet. Dies betrifft nicht nur das Gesamtlexikon, das um neue Einträge erweitert werden kann, sondern auch die einzelnen Einträge, die unvollständig sein können.

Auch die zu analysierende Eingabe unterliegt bestimmten Einschränkungen. Der größtmögliche Kontext, der für die Untersuchung eines unbekannten Wortes herangezogen wird, ist hier der Satz. Über Satzgrenzen hinausgehende Eingaben (z.B.

---

[1] Andere Arbeiten, die sich mit der Simulation lexikalischen Erwerbs beschäftigen, finden sich vor allem in Zernik 89. Die dort präsentierten Arbeiten zeigen das zunehmende theoretische und praktische Interesse an dieser Thematik.

Texte) werden in dieser Arbeit nicht berücksichtigt.

Die prozeduralen Aspekte der kognitiven Verarbeitung werden durch einen left-corner Parser mit top-down Filter modelliert (Wirén 87, Kilbury 90). Dieser Parser simuliert zum einen die Verarbeitung der Eingabekette von links nach rechts. Zum anderen spielen Erwartungen über die als nächste zu analysierende Konstituente eine entscheidende Rolle, was vor allem für die Verarbeitung von unbekannten Wörtern wichtig ist. In einem Durchgang wird zugleich der Eingabesatz analysiert und den unbekannten Wörtern alle erschließbare Information zugewiesen.

Den theoretischen Rahmen dieser Arbeit bildet das Unifikationsparadigma (vgl. Shieber 86). Für die Analyse natürlicher Sprache folgt daraus, daß Eingabeketten entsprechend einem unifikationsbasierten Grammatikfragment verarbeitet werden. Diese Verarbeitung besitzt die für unifikationsbasierte Ansätze charakteristischen Eigenschaften der Monotonie und des inkrementellen Aufbaus von Repräsentationen. Auch die Analyse unbekannter Wörter, die denselben Prinzipien unterliegt, wird von diesen Eigenschaften bestimmt. Dies führt dazu, daß Information über ein unbekanntes Wort ebenfalls inkrementell und monoton gewonnen wird. Mittels Unifikation wird dabei der sprachliche Kontext des Wortes, d.h. der Eingabesatz und die ihm zugeordnete Analyse, als Quelle aller Information genutzt.[2]

Neben diesen Grundannahmen, die den Lösungsansatz bestimmen, ist die Klärung der Frage, in welchen Fällen ein Wort als unbekannt gilt, zentral für die Simulation lexikalischen Erwerbs. Grundsätzlich gilt, daß Wörter dann unbekannt sind, wenn ein passender Lexikoneintrag fehlt. Dieses Fehlen eines passenden Lexikoneintrages kann einerseits daraus resultieren, daß überhaupt kein Eintrag für diese Wortform vorhanden ist. Ein anderer Grund kann darin bestehen, daß lediglich ein unpassender Eintrag vorliegt, dessen Bestimmung von Wortklasse, Subkategorisierung o.ä. dem konkreten Kontext nicht entspricht. Damit ein Wort als unbekannt gilt, muß es aber weiteren Kriterien genügen. So muß das Wort phonetisch und morphologisch zulässig sein, d.h. allgemeine Gesetzmäßigkeiten für den Aufbau von Wörtern der betrachteten Sprache (hier deutsch) dürfen nicht verletzt werden. Dieses Kriterium soll dazu dienen, die unbekannten Wörter von Tippfehlern zu unterscheiden. Weiterhin unterliegt das Wort Einschränkungen, die aus der Unterscheidung von offenen und geschlossenen Wortklassen resultieren. Als geschlossene Wortklassen betrachten wir Funktionswörter wie Konjunktionen, Artikel, Pronomen u.ä., dagegen gelten die

[2] Die Annahme der Monotonie ist jedoch auf die **Analyse** natürlicher Sprache beschränkt. Wie weiter unten in Kap. 3 gezeigt wird, werden für die **Repräsentationen** im Lexikon durchaus nichtmonotone Mittel verwendet. Unabhängig davon muß die Frage beantwortet werden, welche Mittel für die Modifikation bestehender Lexikoneinträge erforderlich sind.

Wortklassen der Substantive, Verben, Adjektive und Adverbien grundsätzlich als offen. Innerhalb dieser offenen Wortklassen gibt es allerdings semantisch bedingte geschlossene Subklassen wie beispielsweise die Bezeichnungen für Himmelsrichtungen. Aus dieser Unterscheidung von offenen und geschlossenen Wortklassen folgt als Restriktion, daß das unbekannte Wort einer offenen Wortklasse angehören muß. Daher sind beispielsweise keine neuen Artikel oder Pronomen möglich. Eine heuristische Einschränkung besteht in unserer Annahme, daß das unbekannte Wort nicht homonym sein kann mit einer Wortform einer geschlossenen Klasse. So kann z.B. die Wortform *der*, für die es bereits einen Lexikoneintrag als definiten Artikel gibt, nicht als ein neues Substantiv analysiert werden. Schließlich gehen wir von der Annahme aus, daß unbekannte Wörter in allen Aspekten regelmäßig sind. Dies betrifft beispielsweise ihre morphologischen Eigenschaften oder auch die Subkategorisierung von Verben (u.a. für eine Nominalphrase im Nominativ als Subjekt).

All diese Annahmen und Postulate, Fragestellungen und Restriktionen sind notwendig, um die menschlichen Sprachfähigkeit angemessen zu modellieren und werden daher bei der vorgestellten Simulation lexikalischen Erwerbs beachtet.

2 Gewinn der Information für unbekannte Wörter

Um die Frage, wie die Information für unbekannte Wörter gewonnen wird, zu klären, muß zunächst die generelle Vorgehensweise festgelegt werden. Dazu gehört die Bestimmung eines formalen Rahmens und des angestrebten Ergebnisses, die sich gegenseitig bedingen. Diese generelle Vorgehensweise ist dann in die Architektur eines zu implementierenden Systems umzusetzen. Dieses System (realisiert als das QPATR-System, vgl. Kilbury 90) erlaubt schließlich die Simulation des lexikalischen Erwerbs auf dem Computer.

2.1 Aufsammeln der Information für unbekannte Wörter bei der Analyse

Wie bereits in Kap. 1 erwähnt, bildet die Unifikationgrammatik hier den formalen Rahmen für die Analyse natürlicher Sprache. Dabei verwenden wir die Unifikationsgrammatik in ihrer einfachsten Ausprägung, d.h. der einzig zulässige Datentyp in diesem Formalismus ist die Merkmalsstruktur und die einzig mögliche Operation auf diesem Datentyp ist die Unifikation. Jedem sprachlichen Zeichen, d.h. hier sowohl einfachen lexikalischen Wortformen als auch komplexen phrasalen

Konstituenten und Sätzen, werden Merkmalsstrukturen zugeordnet, die die Information, die mit diesem Zeichen verbunden ist, repräsentieren.

Jede Merkmalsstruktur kann durch Unifikation mit einer anderen (kompatiblen und nichtgleichen) Merkmalsstruktur inkrementell erweitert werden. Merkmalsstrukturen sind gut geeignet zur Repräsentation von unterspezifizierter Information, und sie können Informationen aller linguistischen Ebenen repräsentieren (d.h. morphologische, syntaktische, semantische u.a.).

Neben den Merkmalsstrukturen, die den sprachlichen Zeichen zugeordnet sind, gibt es auch solche, die in Grammatikregeln vorkommen. Die Grammatikregeln in einer Unifikationsgrammatik bestehen aus einem kontextfreien Teil, der die sprachlichen Zeichen mittels Konkatenation verkettet und Merkmalsstrukturen, die festlegen, welche Teile der mit den sprachlichen Zeichen verbundenen Merkmalsstrukturen unifiziert werden. In den Merkmalsstrukturen der Regeln können generelle grammatische Zusammenhänge wie beispielsweise die Weitergabe von Kopf- oder Kongruenzmerkmalen durch die Verwendung von Templates ausgedrückt werden.[3]

Die Repräsentation der einzelnen lexikalischen Zeichen und die Festlegungen in den Regeln führen dazu, daß auch komplexen Phrasen Merkmalsstrukturen zugeordnet werden. Da die Unifikation eine nichtdirektionale Operation ist, folgt außerdem aus dem monotonen und inkrementellen Informationsaufbau, daß auch Merkmalsstrukturen einzelner lexikalischer Zeichen durch ihre Unifikation mit denen der Grammatikregeln zusätzliche Informationen erhalten können. Auf diese Weise wird den unbekannten Wörtern aufgrund des sprachlichen Kontextes alle relevante Information, repräsentiert in Merkmalsstrukturen, zugewiesen. Diese Informationszuordnung erweist sich als völlig ausreichend, um entsprechende neue Lexikoneinträge zu erstellen.

Während die Information selbst aus dem Lexikon und den Regeln stammt (gemäß den bekannten Wörtern des sprachlichen Kontextes), ist es der Parser, der die Verarbeitung steuert. Erst die Interaktion von Parser, Regeln und Lexikon ermöglicht eine angemessene Analyse und das Aufsammeln der Information für unbekannte Wörter.

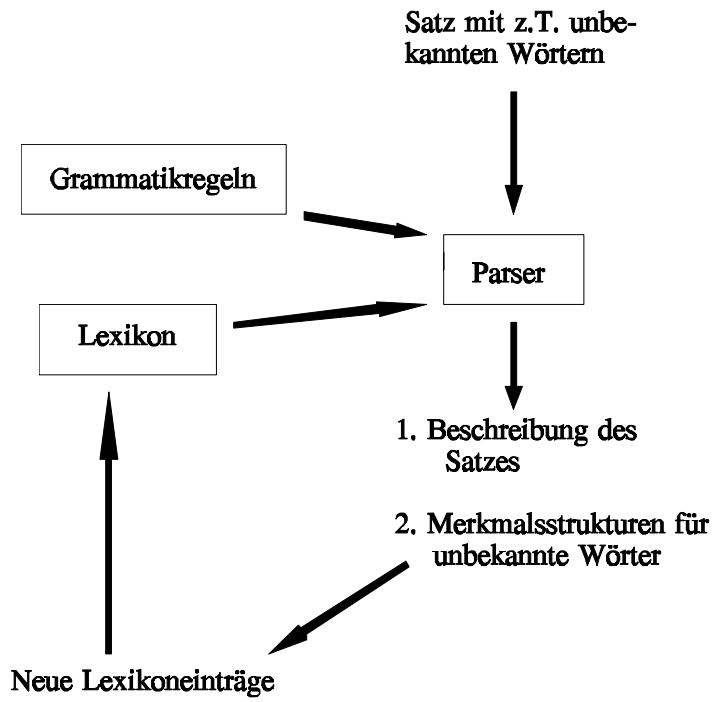
---

[3] In manchen unifikationsbasierten Theorien wie beispielsweise der Head-Driven Phrase Structure Grammar - HPSG (vgl. Pollard/Sag 87) werden solche allgemeinen Zusammenhänge durch Prinzipien ausgedrückt, die fester Bestandteil des Formalismus sind. In werkzeugorientierten Unifikationsformalismen wie PATR-II und auch QPATR sind solche Mittel nicht vorgesehen. Die Verwendung eines werkzeugorientierten Formalismus legt daher keine spezielle Grammatiktheorie fest und sichert die Möglichkeit, Ergebnisse aus verschiedenen Theorien zu übertragen.

2.2 Architektur von QPATR

Folgende Abbildung illustriert die Umsetzung der oben angeführten Überlegungen in eine konkrete Systemarchitektur.

(1) Systemarchitektur von QPATR



Wie (1) zeigt, wird dem Parser ein Satz, der auch dem System unbekannte Wörter enthalten kann, eingegeben. Dieser Parser muß verschiedene Kriterien erfüllen. So muß es sich um einen robusten Parser handeln, in dem Sinne, daß er bei unbekannten Wörtern nicht einfach scheitert. Statt dessen soll er aufgrund der Grammatikregeln und der anderen Wörter des Satzes, die ihre Information aus dem Lexikon erhalten, Analysen sowohl des Satzes als auch des unbekannten Wortes (und entsprechend der Analyse des Satzes) vorschlagen. Außerdem handelt es sich um einen grammatikunabhängigen Parser, der beliebige Grammatikfragmente und beliebige Sprachausschnitte verarbeiten kann, da er nur den Algorithmus (hier: left-corner mit linking-Relation) umsetzt und die Verarbeitung des Eingabesatzes auf den Grammatikregeln und dem Lexikon beruht. Da für die unbekannten Wörter ein einfaches Suchen im Lexikon nicht ausreicht - charakteristisch für unbekannte Wörter ist, daß es keinen angemessenen Lexikoneintrag gibt -, muß diese Komponente des Parsers modifiziert

werden. Mit Hilfe der linking-Relation werden Erwartungen über die als nächste zu analysierende Konstituente gewonnen. Wenn bei einer lexikalischen Konstituente die Suche im Lexikon scheitert, handelt es sich um ein unbekanntes Wort (sofern die anderen Bedingungen auch alle erfüllt sind). Ausgehend von der Erwartung wird eine Hypothese gebildet und die Repräsentation der Erwartung (in Form einer Merkmalsstruktur) als unterspezifizierte Repräsentation dem unbekannten Wort zugeordnet. Aufgrund der weiteren Analyse des Satzes kann diese Repräsentation weiterspezifiziert werden. Die Modifikation des Parsers betrifft somit nur die Komponente, die die Suche im Lexikon leistet.

Da es sich hier um die Implementation eines allgemeinen Parsing-Algorithmus handelt, wird für die Analyse natürlicher Sprache ein vom Parser unabhängiges Grammatikfragment benötigt, in dem ein konkreter Ausschnitt einer gewählten Einzelsprache beschrieben wird. Dieses Fragment besteht zum einen aus Grammatikregeln. Wie bereits erwähnt, handelt es sich um kontextfreie Phrasenstrukturregeln, deren kontextfreier Teil um Merkmalsstrukturen erweitert ist (vgl. Shieber 86).

Neben den Grammatikregeln umfaßt das Grammatikfragment ein Lexikon. Entsprechend einer der Grundannahmen betrachten wir das Lexikon sowohl insgesamt als auch in den einzelnen Einträgen als unvollständig und prinzipiell erweiterbar. Innerhalb der QPATR-Architektur ist es die grundlegende Aufgabe des Lexikons[4], den lexikalischen Zeichen Merkmalsstrukturen als Repräsentationen zuzuordnen, die dann entsprechend den Grammatikregeln bei der Analyse eines Eingabesatzes verarbeitet werden.

Ausgehend von dem Eingabesatz mit z. T. unbekannten Wörtern, den Grammatikregeln und dem Lexikon leistet der Parser zweierlei. Erstens liefert er, wie jeder Parser eines sprachverarbeitenden Systems, eine Beschreibung des Eingabesatzes. Da es sich hier um einen unifikationsbasierten Formalismus handelt, hat die Beschreibung die Form einer Merkmalsstruktur. Zweitens, und dies unterscheidet die Architektur des QPATR-Systems von anderen sprachverarbeitenden Systemen, wird auch für unbekannte Wörter eine Beschreibung geliefert.

Die Beschreibung für ein unbekanntes Wort ist ebenfalls eine Merkmalsstruktur. Obwohl diese Merkmalsstruktur alle dem Wort zugehörige Information repräsentiert, ist sie trotzdem kein geeigneter Lexikoneintrag. Denn ein Lexikoneintrag

---

[4] In den neueren unifikationsbasierten Arbeiten kommt dem Lexikon eine immer stärkere Bedeutung zu. Aufgrund der lexikalistischen Ausrichtung (z.B. HPSG, Unification Categorial Grammar - UCG - Zeevat/Klein/Calder 86) wird immer weniger Information in Regeln und immer mehr Information im Lexikon repräsentiert, wodurch das Lexikon zur zentralen Komponente eines sprachverarbeitenden Systems wird.

dient nicht nur dazu, eine bloße Informationsansammlung für die Analyse zu sein, sondern er muß auch die Beziehungen und Zusammenhänge, die zwischen lexikalischen Zeichen bestehen, abbilden. Diese Abbildung kann nicht von Merkmalsstrukturen geleistet werden, weshalb sie als Lexikoneinträge ausscheiden.

In unserem System haben Lexikoneinträge eine wohldefinierte Form gemäß unseren Überlegungen zu einem Lexikonmodell. Aufgabe des Systems ist nun, aus der Merkmalsstruktur, die dem unbekannten Wort zugeordnet ist, einen angemessenen Lexikoneintrag für dieses unbekannte Wort zu formulieren. Indem der neue Lexikoneintrag die vorgegebene Struktur des Lexikons und der Lexikoneinträge beachtet, wird eine Integration in das Lexikon erreicht.

3 Modell des Lexikons

Das Lexikon ist die zentrale Komponente beim lexikalischen Erwerb, da einerseits Informationen für unbekannte Wörter aus den Lexikoneinträgen für bekannte Wörter erschlossen werden und andererseits neue Lexikoneinträge für unbekannte Wörter in das bestehende Lexikon integriert werden. Um diese Aufgaben zu erfüllen, muß ein Modell erstellt werden, das die generellen Anforderungen an das Lexikon, seine Struktur und seine Eigenschaften bestimmt. Wie schon in Kap. 1 erwähnt, ist die Unvollständigkeit des Lexikons eine charakteristische Eigenschaft, die die prinzipielle Erweiterbarkeit zur Folge hat. Da neue Wörter aufgrund von Wortbildung (Komposition, Derivation), aber auch aufgrund von Übernahmen aus anderen Sprachen (Fremd-, Lehnwörter) lexikalisiert werden können und bekannte Wörter aufgrund von Sprachwandel ihre Eigenschaften (v.a. semantische, aber auch syntaktische und morphologische) verändern können, muß ein Modell des Lexikons diese Eigenschaft der Unvollständigkeit zusammen mit ihren Folgen in den Mittelpunkt der Simulation stellen.

Das Lexikon enthält alle Arten von linguistischer Information. Zwar gelten für die jeweiligen Arten von Information eigene Prinzipien (semantische Prinzipien für semantische Merkmale, syntaktische Prinzipien für syntaktische Merkmale), doch muß in jedem lexikalischen Zeichen die für dieses Zeichen relevante Information (morphologisch, syntaktisch u.a.) gemeinsam spezifiziert sein.

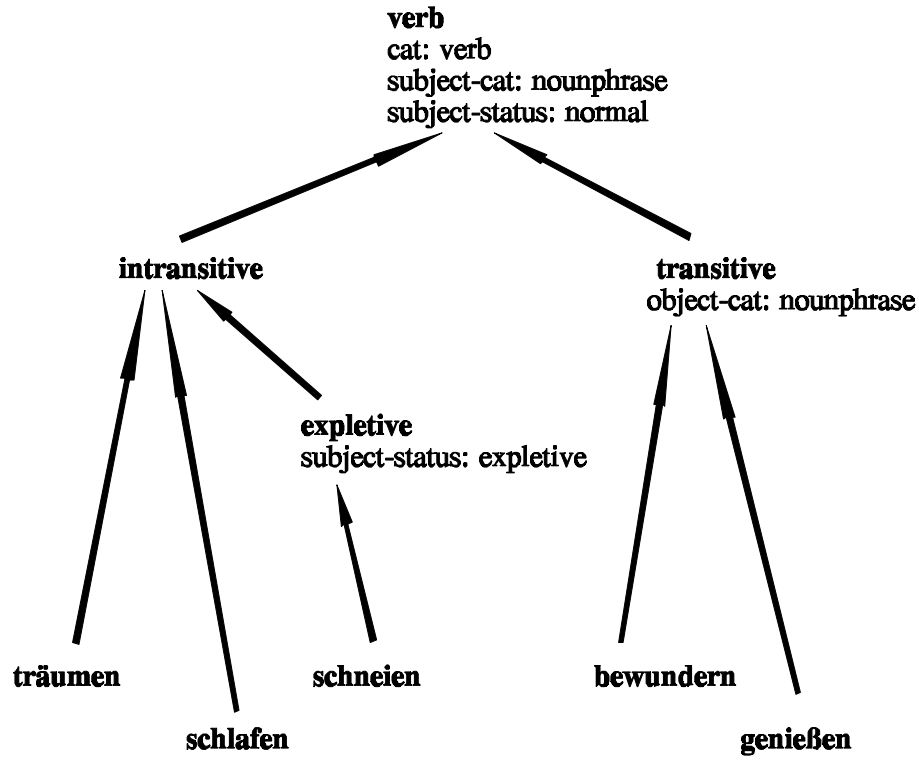
Zugleich muß ein Lexikon als Teil eines implementierten Systems ein Format aufweisen, das nicht nur für das System geeignet ist, sondern auch für die Benutzer des Systems leserlich ist. Die in der Arbeit mit Computern übliche Anforderung an eine benutzerfreundliche Bedienung gilt hier somit entsprechend in der Form, daß die

die im Lexikon verwendete Notation verständlich sein soll.

Ein adäquates Lexikonmodell erfordert einfache Repräsentationen mit minimaler Redundanz (vgl. Kiparsky 82). Wird die gleiche Information an unterschiedlichen Stellen verwendet, sollte sie nicht jedesmal definiert werden, sondern durch Verweise auf eine einzige Definition repräsentiert werden. Diese Verweise dienen aber nicht nur dazu, Redundanz zu vermindern, sondern sollen auch zu einer linguistisch adäquaten Strukturierung der Information führen. Zusammenhänge und Beziehungen, die zwischen lexikalischen Zeichen bestehen, können durch eine solche Verweisstruktur abgebildet werden. Diese Darstellung von Information führt zu einem strukturierten Lexikon (vgl. Flickinger/Pollard/Wasow 85, Shieber 86, Pollard/Sag 87).

Die Beziehungen, die im Lexikon gelten, können als Klassen- oder Typen-Zugehörigkeit bezeichnet werden. Die Annahme von abstrakten linguistischen Typen (wie Substantiv, intransitives Verb, Adjektiv u.ä.), die den Begriff der traditionellen Wortklasse ersetzen, ist dabei notwendig. Diese Typen spezifizieren all die Informationen, die den einzelnen lexikalischen Zeichen oder Subtypen gemeinsam sind. Einfache Verweise auf einen Typ reichen zwar aus, um diese Information weiterzugeben, sie sind aber für die folgende Besonderheit der Informationsstrukturierung im Lexikon unzureichend. Neben den lexikalischen Zeichen und Subtypen, die alle Information ihres übergeordneten Typs aufweisen, gibt es auch solche, die die meiste Information eines bestimmten Typs besitzen, sich aber in einzelnen Eigenschaften von diesem Typ unterscheiden. Für die Beziehungen im Lexikon gilt daher, daß sie nicht nur regelmäßige Beziehungen, sondern auch Ausnahmen und Unregelmäßigkeiten erfassen müssen.

(2) Beziehungen im Lexikon



In dem Beispiel in (2) wird ein allgemeiner Typ **verb** angenommen, der die für Verben typischen Eigenschaften besitzt (Kategoriebezeichnung Verb, eine Nominalphrase als Subjekt mit Status normal, d.h. nichtexpletiv). Subklassen dieses Typs sind **intransitive** und **transitive**, wobei sich die transitiven Verben von dem übergeordneten Typ dadurch unterscheiden, daß sie zusätzliche Information über ein Objekt besitzen. Lexikalische Zeichen wie die Wortformen *bewundern* und *genießen* sind solche transitiven Verben. Auch auf die intransitiven Verben verweisen lexikalische Zeichen (*träumen*, *schlafen*). Zugleich verweist aber auch ein Subtyp auf **intransitive**: der Subtyp **expletive**. Dieser Subtyp ist eine Unterklasse der intransitiven Verben. Wie sie regieren diese Verben nur ein Subjekt, doch abweichend von ihrem übergeordneten Typ regieren sie ein expletives Subjekt (subject-status: expletive). Alle anderen, nicht abweichenden Eigenschaften weist ihnen ihr übergeordneter Typ zu, doch diese spezifische Eigenschaft "überschreibt" die entsprechende Eigenschaft des übergeordneten Typs. Ein solches expletives Verb, das alle Eigenschaften von **expletive** aufweist, ist *schneien*.

Für die Modellierung des Lexikons ist es notwendig, daß diese Art von Beziehungen abgebildet werden kann. Während für den generellen Verweis eine Informationsweitergabe mittels Unifikation ausreichen würde, kann das Überschreiben von

Information, wofür nichtmonotone Mittel erforderlich sind, nicht in den (üblicherweise monotonen) unifikationsbasierten Formalismen ausgedrückt werden.[5] Da aber gerade eine Beziehung wie "X ist ein Y ausgenommen die Eigenschaft Z" für das Lexikon charakteristisch ist, benötigen wir für die Repräsentation des Lexikons einen Formalismus, für den diese Beziehung im Mittelpunkt steht. Eine Möglichkeit, eine solche Beziehung zu repräsentieren, bildet die Default-Vererbung. Mit ihr kann ausgedrückt werden, daß Eigenschaften einer übergeordneten Klasse geerbt werden (als Standardannahmen bzw. Defaults), wenn keine spezifische Angabe gemacht wird.

Eine Repräsentationssprache, die über diese Ausdrucksmittel verfügt, ist DATR, entwickelt von Roger Evans und Gerald Gazdar (Evans/Gazdar 89a, 89b, 90). DATR ist eine deklarative Sprache mit expliziter Inferenztheorie zur Repräsentation einer eingeschränkten Klasse von semantischen Netzen, die sowohl Mehrfachvererbung als auch Default-Vererbung zulassen.

Die mittels DATR beschriebenen semantischen Netze werden Theorien genannt und haben ein festgelegtes Format. Jede Theorie besteht aus einem oder mehreren Knoten, beginnend mit einem Großbuchstaben. Jeder Knoten wird durch Pfad-Wert-Paare definiert. Pfade bestehen aus beliebig vielen Attributen, die von spitzen Klammern umschlossen sind, wobei der Pfad auch kein Attribut enthalten kann (leerer Pfad). Werte können entweder beim Pfad selbst spezifiziert sein (atomar oder Listenwert) oder mittels Verweis von einem anderen Knoten, Pfad oder Knoten-Pfad-Paar geerbt werden. Das folgende Beispiel mit den beiden DATR-Knoten **Knoten1** und **Knoten2** soll die Syntax von DATR-Theorien verdeutlichen:

(3) einfache DATR-Theorie

```
Knoten1:    <attribut1 attribut2> == atomarer_wert
            <attribut3> == <attribut1 attribut2>.
Knoten2:    <> == Knoten1:<attribut1 attribut2>
            <attribut1 attribut2> == Knoten1.
```

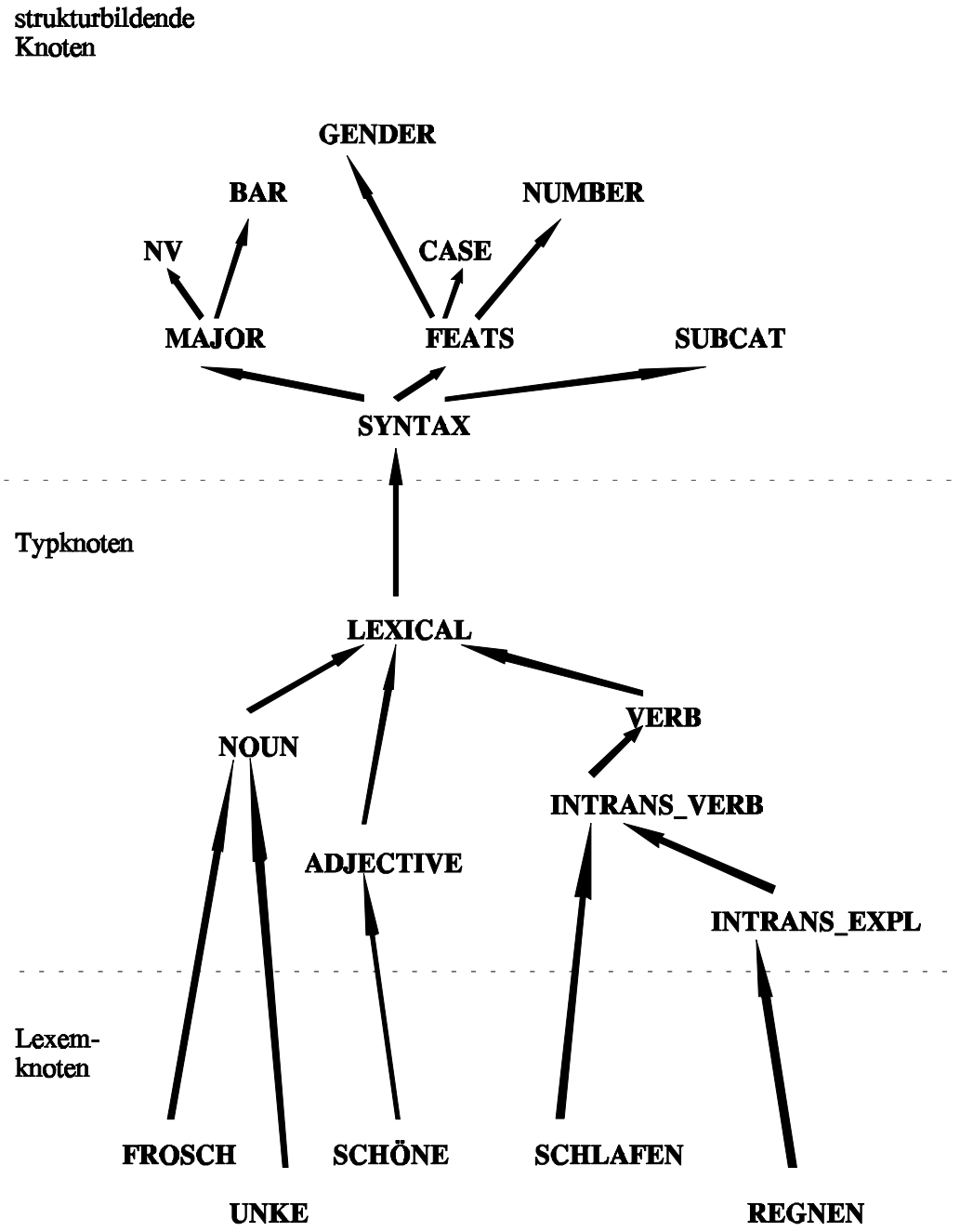
Zugang zu der Information, die in den DATR-Theorien repräsentiert ist, wird durch Abfragen gefunden. Abfragen bestehen aus einem Knoten-Pfad-Paar, für deren Evaluierung in der DATR-Theorie die entsprechende Knotendefinition herangezogen wird. Enthält die Knotendefinition den Abfragepfad, wird die rechte Seite des betreffenden DATR-Satzes mittels einer von sieben Inferenzregeln evaluiert, d.h. sie wird

[5] Eine wohldefinierte nichtmonotone Erweiterung für unifikationsbasierte Formalismen wird jedoch in Bouma 90 vorgeschlagen.

zur neuen Abfrage. Ist der Abfragepfad nicht Teil der Knotendefinition, kommt der Defaultmechanismus zum Tragen. In diesem Fall wird per default derjenige Pfad der Knotendefinition gewählt, der die längste passende Anfangssequenz (Präfix) des Abfragepfades darstellt. So würde im Beispiel (3) für alle Abfragen bezüglich **Knoten2**, deren Pfad nicht mit der Attributfolge **<attribut1 attribut2>** beginnt, der leere Pfad **<>** gewählt, der Präfix aller Pfade ist, und dessen rechte Seite evaluiert. Wenn es keinen passenden Pfad gibt (z.B. für die Abfrage **Knoten1:<attribut2>**), scheitert die Evaluierung. Alle sinnvollen Abfragen für die Theorie in (3) führen zu dem atomaren Wert **atomarer_wert**.

Obwohl DATR als lexikalische Repräsentationssprache entwickelt wurde und häufig für diese Aufgabe eingesetzt wird (vgl. Cahill/Evans 90, Kilbury/Naerger/Renz 91, Gibbon 92), ist es ein werkzeugorientierter Formalismus. Unterschiedliche Annahmen und Eigenschaften, abhängig von dem zugrundeliegenden Modell des Lexikons, können in DATR repräsentiert werden. Dies bedeutet, daß auch das Konzept von Typen, auf die andere Subtypen und lexikalische Zeichen verweisen und die die gemeinsamen Eigenschaften angeben, in DATR realisiert werden kann. Dieses Konzept wird in Abbildung (4) beispielhaft illustriert.

(4) Struktur des Lexikons



Die in (4) vorgeschlagene Struktur des Lexikons unterscheidet drei Ebenen. Die erste Ebene bilden die strukturbildenden Knoten. Sie definieren Teile der Merk-

malsstrukturen in der Form, die vom QPATR-System benötigt wird.[6] Zugleich typisieren sie diese Merkmalsstrukturen, indem sie zum einen die vorkommenden Attribute und zum anderen den Wertebereich eines jeden Attributs festlegen. So wird beispielsweise definiert, daß jedes Substantiv für Numerus spezifiziert ist, wobei die möglichen Werte singular und plural sind. Dagegen ist ausgeschlossen, daß die Beschreibung einer Präposition ein Merkmal für Numerus aufweist, oder daß ein Substantiv einen anderen Numeruswert als singular oder plural erhält.[7]

Die zweite Ebene bilden die in dem vorgestellten Lexikonmodell zentralen Typknoten, die generelle lexikalische Information zusammenfassen. Eine Teilmenge der Typknoten bilden die lexikalischen Typknoten, die diejenigen sind, auf die die Lexemknoten direkt verweisen. Alle Typknoten sind hierarchisch nach Generalität geordnet. Da jeder Typknoten auf genau einen generelleren Knoten verweist, ergibt sich für diese Ebene des Lexikons eine Baumstruktur. Die Evaluierung der Typknoten führt zu unterspezifizierten Merkmalsstrukturen, die all die Information enthalten, die typisch ist für Knoten, die auf diesen Typknoten verweisen. Zu beachten ist hier, daß diese typische Information aufgrund der Verwendung von Default-Vererbung durch spezifischere Information überschrieben werden kann. An dieser Stelle werden die subregulären und irregulären Beziehungen, die für das Lexikon charakteristisch sind, modelliert.

Die dritte Ebene bilden die Lexemknoten. Sie stellen die eigentlichen Lexikoneinträge dar. Jeder Lexemknoten verweist auf genau einen lexikalischen Typ. Dies spiegelt die Annahme wider, daß die Lexemklassen disjunkt sind. Außer dem Verweis auf ihren Typknoten enthalten die Lexemknoten die lexemspezifische Information. Diese lexemspezifische Information ergänzt die des Typknotens, überschreibt sie jedoch nicht. Wie sich später zeigen wird, ist diese Annahme, daß Irregularitäten in Typknoten und nicht in Lexemknoten residieren, notwendig für die Erstellung neuer Lexikoneinträge. Auch die Lexemknoten sind evaluierbar. Ihre Evaluation führt zu vollspezifizierten Merkmalsstrukturen, die im QPATR-System den lexikalischen Zeichen zugeordnet werden.

---

[6] Zwar ähnelt die Syntax von DATR der von PATR-II, doch unterscheiden sich die beiden Formalismen grundsätzlich in ihrer Semantik. Da das QPATR-System auf Merkmalsstrukturen operiert und Merkmalsstrukturen kein Bestandteil von DATR sind, muß eine Verbindung zwischen den beiden Formalismen geschaffen werden, um DATR zur Repräsentation des QPATR-Lexikons zu verwenden. Diese Verbindung kann darin bestehen, daß die Evaluierung von DATR-Knoten zu einer metasprachlichen Beschreibung von Merkmalsstrukturen führt. Wie in Kilbury/Naerger/Renz 91 erläutert, ist auf diese Weise keine zusätzliche Schnittstelle notwendig.

[7] In diesem Sinne sind unsere Typen geschlossen wie die in der UCG (vgl. Moens et al. 89). Dadurch unterscheiden sie sich von denen der HPSG, die offen sind.

Beispiele für Lexikoneinträge (die Lexemknoten **FROSCH** und **SCHÖNES**) und ihre zugehörigen lexikalischen Typen (die Typknoten **NOUN** und **ADJECTIVE**) sind in (5) aufgelistet.[8]

(5) Lexikoneinträge und lexikalische Typen

FROSCH: <> == NOUN
<gender> == masculine
<number> == singular
<case> == nominative.

SCHÖNES: <> == ADJECTIVE
<gender> == neuter
<number> == singular
<case> == nominative
<inflection> == strong.

NOUN: <> == LEXICAL
<cat> == noun
<person> == third.

ADJECTIVE: <> == LEXICAL
<cat> == adjective.

Eine derartige Strukturierung der Lexikoneinträge erlaubt die linguistisch sinnvolle Unterscheidung von zwei Arten von Information: zum einen die lexemspezifische Information, die direkt in den Lexemknoten durch Pfade mit atomaren Werten ausgedrückt ist, und zum anderen die allgemeine, typische Information, die der Lexemknoten mit den anderen Lexemknoten dieses Typs gemeinsam hat. Diese Typinformation wird durch den leeren Pfad vom lexikalischen Typ ererbt.

Jede Erweiterung des bestehenden Lexikons muß die vorgegebene Struktur (sowohl die des Gesamtlexikons als auch die der einzelnen Einträge) beachten, da diese Struktur zentraler Bestandteil unseres Lexikonmodells ist.

Zwei Arten der Erweiterung des Lexikons müssen unterschieden werden. Wenn die Information eines vorhandenen Lexikoneintrages zu der Verwendung der gleich-

---

[8] In den folgenden Beispielen werden zur vereinfachten Darstellung Vollformen verwendet. Dies zeigt sich sowohl an der Form der Lexemknoten als auch an der repräsentierten Information. Außerdem wird die generelle Vorgehensweise ausschließlich an morphologischen und syntaktischen Informationen illustriert, Informationen anderer linguistischer Ebenen (z.B. semantische) können auf die gleiche Weise behandelt werden.

lautenden Form in einem konkreten Kontext im Widerspruch steht, muß dieser Lexikoneintrag überarbeitet werden.[9] Im anderen Fall, wenn kein Lexikoneintrag existiert, muß ein völlig neuer Lexikoneintrag formuliert werden, indem ein neuer Lexemknoten mit der für ihn geltenden Information gebildet wird. Während die Erweiterung von bestehenden Lexikoneinträgen hier nicht weiter betrachtet werden soll, wird im folgenden gezeigt, wie neue Lexikoneinträge formuliert und in das Gesamtlexikon integriert werden können. Dabei erweist sich das vorgeschlagene Modell des Lexikons als grundlegende Voraussetzung.

4 Formulierung neuer Lexikoneinträge

Die Formulierung neuer Lexikoneinträge baut zum einen auf unserem Modell des Lexikons, zum anderen auf den Merkmalsstrukturen mit aller relevanten Information für das unbekannte Wort auf. Dabei sind die allgemeine Struktur der Lexikoneinträge und das Konzept der lexikalischen Typen die Bestandteile des Lexikonmodells, die eine entscheidende Rolle für die Erstellung neuer Lexikoneinträge spielen. Denn auch ein neuzubildender Lexikoneintrag muß die kanonische Form von Lexikoneinträgen aufweisen, die aus einem Verweis auf einen generelleren, lexikalischen Typknoten und Pfaden mit den lexemspezifischen Informationen besteht. Das Ziel ist nun, ausgehend von der Merkmalsstruktur, einen solchen Lexikoneintrag zu bilden.

Die Vorgehensweise, um dieses Ziel zu erreichen, gliedert sich in drei Schritte. Zuerst muß der lexikalische Typ, dem das Wort angehört, bestimmt werden. Entsprechend dem Konzept der offenen und geschlossenen Wortklassen nehmen wir ebenfalls eine begrenzte Menge an offenen lexikalischen Typknoten an. Der gesuchte lexikalische Typknoten ist dann derjenige, der dieser Menge angehört und dessen evaluierter Wert (eine unterspezifizierte Merkmalsstruktur) die Merkmalsstruktur des unbekannten Wortes subsumiert. Die Verwendung der Subsumtion zur Bestimmung des lexikalischen Typs ist nur aufgrund der in Kap. 3 aufgestellten Annahme möglich, daß Lexemknoten keine Information ihres Typknotens überschreiben. Ohne diese Annahme könnten die zu vergleichenden Merkmalsstrukturen inkompatible Informationen enthalten, wodurch die Subsumtion scheitern würde.

---

[9] Die Überarbeitung eines bestehenden Lexikoneintrages kann im Extremfall auch darin bestehen, einen zusätzlichen Eintrag zu bilden. Hier muß die Frage nach Homonymie und Polysemie, d.h. ob mehrere Einträge zu einer Form gebildet werden oder disjunkte Information einem Eintrag zugeordnet wird, beantwortet werden. Jede dieser Möglichkeiten kann in dem vorgeschlagenen Ansatz erfaßt werden.

Eine einfache Möglichkeit, den Typknoten zu finden, besteht in der sukzessiven Evaluierung der offenen lexikalischen Typknoten, bis der passende gefunden ist. Eine solche ungerichtete Suche ist allerdings unbefriedigend, wenn die Menge dieser Knoten groß ist. Hier bietet sich eine gerichtete Suche an, die die Typ-Hierarchie verwendet und ausgehend vom generellsten Typ rekursiv zu dem passenden Tochterknoten absteigt, bis der spezifischste passende gefunden ist.

Die folgenden Abbildungen illustrieren in (6) eine Merkmalsstruktur für ein angenommenes unbekanntes Substantiv *Nolf*, das als Subjekt in dem Satz *das Nolf schläft* verwendet wird und aus diesem Satz seine Information erhält, in (7) die Merkmalsstruktur des zugehörigen Typs **NOUN**.

(6) Merkmalsstruktur für ein unbekanntes Substantiv *Nolf*

$$\begin{bmatrix} \textit{syntax}: & \begin{bmatrix} \textit{category}: & \textit{noun} \\ \textit{features}: & \begin{bmatrix} \textit{person}: & \textit{third} \\ \textit{number}: & \textit{singular} \\ \textit{gender}: & \textit{neuter} \\ \textit{case}: & \textit{nominative} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

(7) Merkmalsstruktur des lexikalischen Typs **NOUN**[10]

$$\begin{bmatrix} \textit{syntax}: & \begin{bmatrix} \textit{category}: & \textit{noun} \\ \textit{features}: & \begin{bmatrix} \textit{person}: & \textit{third} \\ \textit{gender}: & [\,] \\ \textit{case}: & [\,] \\ \textit{number}: & [\,] \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

Wenn der passende lexikalische Typknoten gefunden ist, kann er in einem zweiten Schritt dazu verwendet werden, die lexemspezifische Information in der Merkmalsstruktur des unbekannten Wortes zu bestimmen. Diese lexemspezifische Information ist genau die Information, die in der Merkmalsstruktur des unbekannten Wortes

[10] *[ ]* symbolisiert die maximal unterspezifizierte Merkmalsstruktur.

spezifischer ist als in der Typ-Merkmalsstruktur. Um diese Differenzinformation zu finden, wurde ein Algorithmus entwickelt und implementiert, der die beiden Merkmalsstrukturen vergleicht und die gewünschte Information liefert.

Wird dieser Algorithmus auf die Merkmalsstrukturen in (6) und (7) angewendet, ergibt sich die in (8) abgebildete Merkmalsstruktur, die die lexemspezifische Information des Substantives in (6) enthält.

(8) Differenzmerkmalsstruktur der Merkmalsstrukturen in (6) und (7)

$$\begin{bmatrix} \textit{syntax}: & \begin{bmatrix} \textit{features}: & \begin{bmatrix} \textit{number}: & \textit{singular} \\ \textit{gender}: & \textit{neuter} \\ \textit{case}: & \textit{nominative} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

Der letzte Schritt besteht nun darin, ausgehend von dem gefundenen lexikalischen Typknoten und der lexemspezifischen Information einen neuen Lexikoneintrag zu formulieren. Aus der Orthographie des unbekannten Wortes wird der neue Lexemknoten gebildet, dessen leerer Pfad auf den lexikalischen Typknoten verweist, und dessen lexemspezifische Information durch Pfad-Wert-Paare ausgedrückt wird. Die Ermittlung dieser Pfad-Wert-Paare basiert auf einer Entsprechung in unserem Lexikon zwischen den PATR-Pfaden einer Merkmalsstruktur und den Pfad-Wert-Paaren der DATR-Theorie. Im allgemeinen entspricht das letzte Attribut und dessen Wert in einem PATR-Pfad dem Pfad-Wert-Paar in der zugehörigen DATR-Theorie.[11]

Ausgehend von der Merkmalsstruktur in (8) und dem gefundenen lexikalischen Typ kann der folgende Lexikoneintrag für die Wortform *Nolf* formuliert werden.

(9) Neuer Lexikoneintrag für *Nolf*

NOLF: <> == NOUN
<gender> == neuter
<case> == nominative
<number> == singular.

[11] Im Fall der Subkategorisierungsinformation von Verben genügt es nicht, lediglich das letzte Attribut eines PATR-Pfades zu berücksichtigen. Da die Art der Ergänzung (Subjekt, direktes Objekt, indirektes Objekt) von der Position in der Subkategorisierungsliste bestimmt wird, muß der ganze Pfad, der diese Information repräsentiert, berücksichtigt werden.

Durch diese Vorgehensweise entspricht die Struktur des neuen Lexikoneintrages der kanonischen Form aller Lexikoneinträge, und der neue Lexikoneintrag ist aufgrund des Verweises auf den zugehörigen Typknoten in die Netzstruktur des bestehenden Lexikons integriert.

5 Ausblick

Der vorgestellte Lösungsansatz zeigt, wie neue Lexikoneinträge für unbekannte Wörter in einem unifikationsbasierten System formuliert und in das bestehende Lexikon integriert werden können. Auf diese Weise wird ein Aspekt des lexikalischen Erwerbs eines idealisierten erwachsenen Lesers simuliert. Notwendige Voraussetzung für diesen Ansatz ist die Entwicklung eines Lexikonmodells, das als zentrales Konzept lexikalische Typen beinhaltet.

Während die Aufgabe, völlig neue Lexikoneinträge zu erstellen, erfolgreich gelöst wurde, bildet die Erweiterung bestehender Lexikoneinträge um teilweise neue Information einen noch offenen Forschungsgegenstand. Mit einer solchen Erweiterung sind allgemeine Fragen bezüglich der Verwaltung des Lexikons verbunden, deren Beantwortung eine Ergänzung unseres Lexikonmodells darstellen wird.

Literaturverzeichnis

Bouma, Gosse (1990) Defaults in Unification Grammar. In *Proc. of the 28th Conference of the ACL*, 165-171.

Cahill, Lynne J. / Evans, Roger (1990) An Application of DATR: the TIC Lexicon. In *Proceedings of the ECAI-90*, 120-125.

Evans, Roger / Gazdar, Gerald (1989a) Inference in DATR. In *Proc. of the 4th Conference of the EACL*, 66-71.

Evans, Roger / Gazdar, Gerald (1989b) The Semantics of DATR. In A. Cohn (ed.) *AISB89, Proc. of the 7th Conference of the Society for the Study of Artificial Intelligence and Simulation of Behaviour*, 79-87. London: Pitman.

Evans, Roger / Gazdar, Gerald (eds.) (1990) *The DATR Papers: February 1990 (= Cognitive Science Research Paper 139)*. School of Cognitive and Computing Sciences, University of Sussex, Brighton, England.

Flickinger, Daniel / Pollard, Carl / Wasow, Thomas (1985) Structure-Sharing in Lexical Representation. In *Proc. of the 23th Conference of the ACL*, 262-2\20.

Gibbon, Dafydd (1992) ILEX: Linguistic Approach to Computational Lexica. In Ursula Klenk (ed.) *Computatio Linguae*. Beihefte der Zeitschrift für Dialektologie und Linguistik, 32-53.

Kilbury, James (1990) QPATR and Constraint Threading. In *Proc. of 13th COLING*, 382-384.

Kilbury, James / Naerger, Petra / Renz, Ingrid (1991) DATR as a Lexical Component for PATR. In *Proc. of the 5th Conference of the EACL*, 137-142.

Kiparsky, Paul (1982) Lexical Morphology and Phonology. In I.-S. Yang (ed.) *Linguistics in the Morning Calm*, 3-91, Seoul: Hanshin.

Moens, Marc / Calder, Jo / Klein, Ewan / Reape, Mike / Zeevat, Henk (1989) Expressing generalizations in unification-based grammar formalisms. In *Proc. of the 4th Conference of the EACL*, 174-181.

Pollard, Carl / Sag, Ivan (1987) *Information-Based Syntax and Semantics. Vol. I Fundamentals*. Stanford, Calif.: CSLI.

Shieber, Stuart M. (1986) *An Introduction to Unification-Based Approaches to Grammar*. Stanford, Calif.: CSLI.

Wirén, Mats (1987) A Comparison of Rule-Invocation Strategies in Context-Free Chart Parsing. In *Proc. of the 3rd Conference of the EACL*, 226-233.

Zeevat, Henk / Klein, Ewan / Calder, Jo (1986) Unification Categorial Grammar. Centre for Cognitive Science. University of Edinburgh. Edinburgh.

Zernik, Uri (ed.) (1989) *Proc. of the First International Lexical Acquisition Workshop*. Detroit, Michigan.