Rainer Osswald

Die Verwendung von GermaNet zur Pflege und Erweiterung des Computerlexikons HaGenLex

Abstract

Dieser Beitrag soll am Beispiel des semantikbasierten Computerlexikons HaGenLex aufzeigen, wie GermaNet für die Pflege und Erweiterung anderer lexikalisch-semantischer Ressourcen eingesetzt werden kann. Ausgangsbasis ist dabei eine Lesartenzuordnung zwischen GermaNet- und HaGenLex-Einträgen, welche die Übertragung der sinnrelationalen Zusammenhänge von GermaNet auf HaGenLex erlaubt. Auf der Grundlage dieser Kopplung lassen sich beispielsweise Inkonsistenzen in der semantischen Klassifikation von HaGenLex-Einträgen aufdecken. Neben weiteren Anwendungen werden einige sich dabei ergebende Probleme sowie der mögliche Nutzen für die Aufdeckung von Fehlern in GermaNet angesprochen.

1 Das Lexikon HaGenLex

HaGenLex (**Ha**gen **Ge**rma**n** **Lex**icon) ist ein semantikbasiertes Computerlexikon für das Deutsche, das seit 1996 an der FernUniversität Hagen am Lehrgebiet Praktische Informatik VII entwickelt wird. Momentan umfasst es circa 20.000 Lesart-Einträge (etwa 9.200 Substantive, 6.500 Verben und 3.000 Adjektive). Die Einträge wurden primär auf der Grundlage von Frequenzlisten erstellt, mit Unterstützung diverser Wörterbücher des Deutschen. Die Erstellung durch den Lexikographen wird maßgeblich durch eine Werkbank unterstützt, die zum einen die Eingabe leitet, und zum anderen die interne Repräsentation der Einträge als Merkmal-Wert-Strukturen vor dem Nutzer verbirgt bzw. in leicht verständlicher Umschreibung darbietet.

Um Missverständnissen vorzubeugen sei darauf hingewiesen, dass sich HaGenLex von GermaNet in der Gebrauchsweise des Konzeptbegriffs unterscheidet: Während GermaNet, in der Tradition von WordNet, Konzepte durch Synsets repräsentiert sieht, wird in HaGenLex davon ausgegangen, dass jedes lexikalisierte Konzept genau einem Lexem entspricht. Ferner macht HaGenLex, im Gegensatz zu GermaNet, bislang nahezu keinen Gebrauch von künstlichen Konzepten.

Im Folgenden soll der Aufbau von HaGenLex kurz skizziert werden; eine ausführlichere Beschreibung findet sich in (Hartrumpf et al. 2003).

1.1 Der MultiNet-Formalismus

Die Mittel zur Darstellung semantischer Information in HaGenLex sind dem sogenannten *MultiNet-Paradigma* entnommen. Bei diesem handelt es sich um einen Formalismus zur Darstellung der Semantik natürlicher Sprache mittels mehrschichtiger erweiterter semantischer Netze.[1] Grob gesprochen besteht ein solches semantisches Netz aus Knoten, die Konzepte repräsentieren, und Kanten, welche die semantischen Beziehungen zwischen den Konzepten zum Ausdruck bringen.[2] Zur Charakterisierung der Beziehung zwischen Konzepten stellt der MultiNet-Formalismus ein vordefiniertes und ausführlich dokumentiertes Repertoire von weit über hundert Relationen und Funktionen bereit.

Darüber hinaus ist jeder Konzeptknoten von MultiNet hinsichtlich mehrerer Merkmale spezifiziert, die unter anderem zum Ausdruck bringen, ob das Konzept generisch zu interpretieren ist, ob seine Referenz bestimmt oder unbestimmt ist, ob es faktischen oder hypothetischen Charakter hat, und in welcher Weise es einer Quantifikation unterliegt.[3] HaGenLex wurde in

erster Linie zu dem Zweck entwickelt, die automatische Transformation natürlichsprachlicher Ausdrücke in MultiNet-Repräsentationen zu unterstützen.[4] Die hierzu erforderliche syntaktische und semantische Information ist weitgehend lexikalisiert, wobei die Semantik im Lexikon ebenfalls durch MultiNet-Darstellungsmittel geprägt ist. Dazu zählen insbesondere die ontologische Sorte des zugehörigen Konzepts sowie die semantischen Relationen, in denen das Konzept zu anderen Konzepten steht.

1.2 Semantische Klassifikation

Im Rahmen von MultiNet steht eine Hierarchie von 45 ontologischen Sorten zur Klassifikation von Konzepten und damit von Lexemen zur Verfügung. Auf oberster Ebene wird zwischen Objekten, Sachverhalten, Sachverhaltsdeskriptoren, Qualitäten, Graduatoren, Quantitäten und formalen Entitäten unterschieden.

Insbesondere um die Überprüfung von Selektionsrestriktionen zu unterstützen, sind HaGenLex-Lexeme außerdem hinsichtlich 16 binärer semantischer Merkmale klassifiziert. Da zwischen diesen Merkmalen, wie im Fall von HUMAN und ANIMATE, semantische Abhängigkeiten bestehen, sind zulässige Kombinationen von ontologischer Sorte und semantischen Merkmalen zu sogenannten *semantischen Sorten* zusammengefasst. Beispielsweise verbergen sich hinter der semantischen Sorte *con-info* (für 'konkretes Informationsobjekt') die ontologische Sorte *d* (für 'Diskretum') sowie (unter anderem) die semantischen Merkmale [ANIMATE –] (nicht belebt), [ARTIF +] (Artefakt), [INFO +] (Informationsträger) und [MOVABLE +] (beweglich). Subsumierte Lexeme wären in diesem Fall *Abbildung* und *Zeitung*.

1.3 Valenz und Kasusrahmen

HaGenLex spezifiziert die Valenzen von Lexemen sowohl in syntaktischer als auch in semantischer Hinsicht. So ist zu jedem Verb angegeben, in welcher semantischen Beziehung die Partizipanten der vom Verb bezeichneten Situation zu letzterer stehen. Als Ausdrucksmittel finden hierfür wiederum die im Rahmen von MultiNet vorgegebenen semantischen Relationen Verwendung, die insbesondere ein Inventar an thematischen Rollen beinhalten.

In erster Näherung hat der Kasusrahmen für das Verb *informieren* in HaGenLex die folgende Form:

AGT	**OBJ**	**MCONT**
[POTAG +]	[POTAG +]	
np / nom	np / acc	'über'-*pp* / *acc*
	optional	optional

Die erste Zeile listet die thematischen Rollen der Argumente auf, die zweite enthält Selektionsrestriktionen (wobei POTAG für 'potential agent' steht), die dritte gibt (unvollständig) die syntaktischen Valenzen wieder und die letzte Zeile zeigt an, ob es sich um obligatorische oder fakultative Valenzen handelt. Der vollständig spezifizierte Lexikoneintrag ist in Abbildung 2 des Anhangs dargestellt.

1.4 Lexikonstruktur und Datenformat

Die lexikalische Information in HaGenLex ist in Form von *typisierten Merkmal-Wert-Strukturen* repräsentiert. Zugrunde liegt eine baumförmige *Typhierarchie* sowie zu jedem Typ eine *Merkmalsdeklaration*. Die für HaGenLex-Einträge verwendete Merkmal-Wert-Architektur ist im Anhang kurz skizziert.

Zur Strukturierung des Lexikons werden ferner sogenannte *Klassen* eingesetzt, die bestimmte, linguistisch relevante Merkmal-Wert-Kombinationen bündeln. Die Hierarchie der HaGenLex-Klassen ist vererbungsbasiert und erlaubt zudem die Verwendung von Default-Angaben.

2 Vergleich der Darstellungsmittel

Die semantischen Beschreibungsmittel von GermaNet sind darauf ausgerichtet, lexikalisch-se-

mantische Beziehungen zwischen lexikalischen Einheiten bzw. Konzepten darzustellen. Da das MultiNet-Paradigma den Anspruch eines universellen Semantikformalismus erhebt, ist zu erwarten, dass seine Beschreibungsmittel die von GermaNet umfassen.[5]

2.1 Relationen

Die für die Strukturierung von GermaNet und WordNet zentralen, der traditionellen lexikalischen Semantik entliehenen Sinnrelationen werden durch die folgenden MultiNet-Relationen abgedeckt:

SUB	Subordination, Hyponymie
SYNO	Synonymie
ANTO	Antonymie
COMPL	Komplementarität
CONTR	Kontrarität
CNVS	Konversbeziehung

Die drei letztgenannten Relationen sind Unterfälle der Antonymie. Konträr sind zwei Eigenschaften, wenn sie sich gegenseitig ausschließen, wie *heiß* und *kalt*. Komplementäre Eigenschaften müssen sich nicht nur ausschließen, sondern das gesamte Spektrum abdecken; Beispiel: *anwesend* vs. *abwesend*. Die Konversbeziehung zielt dagegen auf Argumentumordnung bei mehrstelligen Prädikaten, wie etwa beim Wechsel zwischen *geben* und *erhalten*. Aus Sicht der lexikalischen Semantik wird man solche Fälle eher unter einen allgemeineren Ansatz zur Behandlung von Diathesen fassen wollen.

Als Ausdrucksmittel vorgesehen sind in GermaNet außerdem eine Kausationsrelation, eine Implikationsbeziehung, eine Teil-Ganzes-Beziehung sowie eine kategorienübergreifende semantische Derivationsrelation (*Pertonymie*).[6] Die folgende Tabelle listet entsprechende, als Substitut taugliche MultiNet-Relationen auf:

CAUS	Kausalbeziehung
IMPL	Implikationsbeziehung
PARS	Teil-Ganzes-Beziehung, Meronymie
CHEA	Wechsel von Ereignis zu Abstraktum

Die letztgenannte Relation CHEA bedarf einer kurzen Erläuterung: Während GermaNet bislang nur eine unspezifische Ableitungsrelation bereitstellt, werden in MultiNet derartige Beziehungen semantisch differenziert ausgedrückt. So bringt CHEA die Beziehung zwischen einem verbalen Ereigniskonzept und dem durch die nominalisierte Form (*nomen actionis*) vermittelten abstrakten Gegenstandskonzept zum Ausdruck, wie etwa zwischen *herstellen* und *Herstellung*. Für weitere derartige „Sortenwechselrelationen" siehe (Helbig 2001).

In diesem Zusammenhang sei angemerkt, dass die semantische Beziehung zwischen einem Verb und seiner Subjekts- bzw. Objektsnominalisierung (*nomen agentis* bzw. *nomen patientis*) in HaGenLex keiner gesonderten semantischen Derivationsrelation bedarf, sondern natürlicherweise durch die entsprechende thematische Rolle des Verbs gegeben ist. Da z.B. im Kasusrahmen von *prüfen* das Subjekt die Rolle AGT (Handelnder) und das direkte Objekt die Rolle OBJ (neutrales Objekt) innehat, ist im Lexikon vermerkt, dass zwischen der Subjektsnominalisierung *Prüfer* und *prüfen* die Relation AGT besteht, und zwischen *Prüfling* und *prüfen* die Relation OBJ.

Abschließend sei noch hervorgehoben, dass der eingangs vermerkte Unterschied zwischen HaGenLex und GermaNet hinsichtlich der Gebrauchsweise des Konzeptbegriffs unter einem formalen Gesichtspunkt vernachlässigbar ist: HaGenLex macht von der Synonymie als *Äquivalenzrelation* Gebrauch, wohingegen GermaNet direkt die daraus resultierenden *Äquivalenzklassen* repräsentiert. Dass sich Hyponymie und Hyperonymie auf ganze Synsets erstrecken, lässt

sich dann durch geeignete MultiNet-Axiome erzwingen:

$\mathrm{SUB}(c_1,c_2) \wedge \mathrm{SYNO}(c_2,c_3) \rightarrow \mathrm{SUB}(c_1,c_3)$,
$\mathrm{SYNO}(c_1,c_2) \wedge \mathrm{SUB}(c_1,c_3) \rightarrow \mathrm{SUB}(c_2,c_3)$.

Entsprechendes gilt für die Reflexivität, Symmetrie und Transitivität der SYNO-Relation.

2.2 Lesartenzuordung

Aufgrund der unterschiedlichen Abdeckung gibt es sowohl HaGenLex-Einträge, die keine GermaNet-Entsprechung haben, als auch den umgekehrten Fall. Momentan beinhalten knapp die Hälfte der HaGenLex-Einträge einen Verweis auf GermaNet-Lesarten (kodiert im lexikalischen Merkmal G-ID; vgl. Anhang).[7]

Bei der Zuordnung der Lesarten eines lexikalischen Wortes lassen sich folgende Fälle unterscheiden:

1. Die Zuordnung ist eineindeutig, d.h. jeder GermaNet-Lesart entspricht genau eine HaGenLex-Lesart, und umgekehrt.
2. Eine GermaNet-Lesart hat keine Entsprechung in HaGenLex.
3. Eine HaGenLex-Lesart hat keine Entsprechung in GermaNet.
4. Eine HaGenLex-Lesart fasst mehrere GermaNet-Lesarten zusammen. Beispielsweise wird *Aal* in GermaNet sowohl als Nahrungsmittel, d.h. als essbare Substanz, als auch als Lebewesen geführt, wohingegen beide Fälle in HaGenLex unter eine Lesart subsumiert sind. (Die diesem Fall zugrunde liegende reguläre Polysemie bzw. Metonymie ist in HaGenLex noch nicht hinreichend erfasst.)
5. Eine GermaNet-Lesart fasst mehrere HaGenLex-Lesarten zusammen. Ein Beispiel hierfür ist der transitive und intransitive Gebrauch von *baden*.

Sei G die Menge der lexikalischen Elemente von GermaNet und H die Menge der mit GermaNet verknüpften HaGenLex-Einträge. Ferner stehe Lgh dafür, dass einer GermaNet-Lesart g die HaGenLex-Lesart h entspricht. (Im Eintrag zu h ist demnach unter dem Merkmal G-ID die Menge $\{ g \in G \mid Lgh \}$ kodiert.) Die auf G gegebene Synonymierelation lässt sich vermittels L auf H projizieren: Sei $\sim$ die kleinste Äquivalenzrelation auf H derart, dass $h \sim h'$ wenn Lgh und $Lg'h'$ für zwei zueinander synonyme Elemente g und g' von G. Mit anderen Worten, $\sim$ ist die transitive Hülle derjenigen symmetrischen und reflexiven Relation auf H, in der zwei Elemente h und h' genau dann stehen, wenn sie GermaNet-Entsprechungen haben, die synonym zueinander sind. (Die Bildung der transitiven Hülle ist aufgrund von Fall 4 erforderlich.)

Die Projektion $\pi(S)$ eines GermaNet-Synsets S auf HaGenLex ist nun folgendermaßen definiert: Steht ein Element von S in der Beziehung L zu einem HaGenLex-Eintrag h, dann sei $\pi(S)$ die zu h gehörige Äquivalenzklasse bzgl. $\sim$; andernfalls sei $\pi(S)$ leer. (Man beachte, dass die Äquivalenzklasse dabei unabhängig von der Wahl von h ist.)

3 Nutzen von GermaNet für HaGenLex

Die Lesarten-Zuordnung erlaubt es, GermaNet in Anwendungen von HaGenLex bei der Informationsrecherche als unterstützende Ressource einzusetzen. Insbesondere können dabei auch GermaNet-Hyperonyme eines in der Anfrage vorkommenden HaGenLex-Lexems herangezogen werden, die keine Entsprechung in HaGenLex haben. Zudem ist über die GermaNet-Verbindung der interlinguale Index von EuroWordNet zugreifbar, wodurch multilinguale Anwendungen, wie die Recherche fremdsprachiger Dokumente mittels deutschsprachiger Anfragen, unterstützt werden.

Im Folgenden wollen wir jedoch nicht die Einsatzmöglichkeiten von GermaNet als ergän-

zende Ressource bei HaGenLex-Anwendungen thematisieren, sondern die Verwendung von GermaNet bei der Pflege und Erweiterung von HaGenLex selbst an zwei Beispielen illustrieren.

3.1 Konsistenzüberprüfung

Eine naheliegende Anwendung der GermaNet-Kopplung besteht darin, die Synset-Projektionen in HaGenLex auf semantische Konsistenz zu überprüfen. Da synonyme Lexeme identische oder zumindest kompatible semantische Sorten aufweisen sollten, ergeben sich dadurch Hinweise auf mögliche Kategorisierungsfehler in HaGenLex. Die auf diese Weise automatisch gewonnenen Hinweise sind dann vom Lexikographen im Einzelnen zu überprüfen. Für eine Inkompatibilität in einer Synset-Projektion kommen als mögliche Fehlerquellen in Betracht:

1. Die semantische Kategorisierung von HaGenLex-Lexemen,
2. die Lesart-Zuordnung zwischen HaGenLex und GermaNet,
3. die Synset-Bildung in GermaNet.

Die ersten beiden Fälle ziehen eine Korrektur von HaGenLex-Einträgen nach sich, wobei der erste Fall eine Qualitätsverbesserung von HaGenLex im engeren Sinne bewirkt.

Der dritte Fall kann auf eine Schwachstelle in GermaNet hindeuten. Das geschilderte Verfahren führt beispielsweise bei der Projektion des GermaNet-Synsets {*Nervosität.1*, *Hektik.1*, ...} zu einer Sorten-Inkompatibilität, da der HaGenLex-Eintrag zu *Nervosität* im Gegensatz zu dem von *Hektik* als (*mentaler*) *Zustand* klassifiziert ist. In der Tat ist die Nervosität einer Person keineswegs mit hektischem Verhalten gleichzusetzen, sodass das Postulat der Synonymie von *Nervosität* und *Hektik* unangemessen scheint. (Man beachte, dass GermaNet keine weiteren Lesarten zu *Nervosität* oder *Hektik* anbietet, und dass {*Gefühl.1, Emotion.1, Empfindung.1, Gemütsbewegung.1*} als Hyperonym fungiert.)

Hinweise auf mögliche Defekte können sich auch aus Fällen ergeben, in denen zwar Kompatibilität gewährleistet ist, aber Unterschiede in der Sortenspezifizität vorliegen. Dies ist etwa bei der Projektion des Synsets {*Nachkomme.1, Kind.1, Nachfahre.1*, ...} der Fall: Die entsprechende Lesart von *Kind* ist in HaGenLex als menschlich klassifiziert, wohingegen *Nachkomme* und *Nachfahre* allgemeiner als Lebewesen eingeordnet sind. Hier ist festzustellen, dass die von GermaNet als synonym zu *Nachkomme* angenommene Lesart von *Kind* zumindest fragwürdig ist. (Als Hyperonym ist übrigens {*Verwandter.1, Angehöriger.1*, ...} vorgesehen.) Zwar ist jedes Kind ein Nachkomme und jeder Nachkomme ist ein Kind von jemandem, um aber als synonym zu gelten, müssten *Nachkomme* und *Kind* in Kontexten wie *Hans ist ein ____ von Adam und Eva* ohne große Bedeutungsverschiebung austauschbar sein, was offenbar nicht der Fall ist. Mit anderen Worten, die Synonymie inhärent relationaler Begriffe muss die zum Ausdruck gebrachte Relation respektieren.

3.2 Restriktion freier Ergänzungen

Neben der Angabe obligatorischer und fakultativer Valenzen schränken Lexeme in HaGenLex auch die möglichen freien Ergänzungen ein, was speziell die Disambiguierung von Präpositionalphrasen unterstützen soll. Zu diesem Zweck sind im Lexikoneintrag diejenigen MultiNet-Relationen aufgelistet, die mit dem zugehörigen Konzept kompatibel sind und damit einer freien Ergänzung semantisch zugrunde liegen können. Beispielsweise sind bei punktuellen Verben wie *platzen* oder *aufwachen* Dauerangaben (ohne iterative Umdeutung) ausgeschlossen, was sich im Lexikoneintrag dadurch niederschlägt, dass die Relation DUR unter dem lexikalischen Merkmal COMPAT-R (kurz für 'compatible relations') nicht aufgeführt ist. Die erforderliche aktionsartliche

resp. aspektuelle Klassifizierung von Verben wird allerdings von der GermaNet-Hierarchie nicht hinreichend unterstützt. So ist *platzen.1*, ebenso wie *reißen.1*, *abkühlen.1* und *trocknen.1*, Hyponym von *?Mat_Zustands_Veränderung.1* ('Veränderung der materiellen Beschaffenheit'); und *aufwachen.1* ist, ebenso wie *altern.1* und *?Mat_Zustands_Veränderung.1*, Hyponym von {*wandeln.3, ändern.1*, ...}. Offenbar geht die Punktualität von Geschehnissen (bzw. deren Ingressivität, Egressivität oder Semelfaktivität) bisher nicht als Strukturierungskriterium in die Verbklassifikation von GermaNet ein.[8]

Im Fall *direktionaler* Ergänzungen, die auf semantischer Ebene durch die MultiNet-Relation DIRCL ('direction/local goal') zum Ausdruck gebracht werden, erscheint dagegen die Heranziehung der GermaNet-Hierarchie auf den ersten Blick vielversprechend. Da die Klasse der Lokationsverben in GermaNet relativ gut strukturiert ist, besteht die Hoffnung, die Klasse der Verben mit möglichen direktionalen Ergänzungen durch eine geeignete Kombination von Einschluss- und Ausschlussklassen herauszufiltern.

Es ergeben sich jedoch verschiedene Schwierigkeiten. Als erstes wäre ein Problem zu nennen, das in der Natur der Sache selbst liegt: Bei Verben, die eine gerichtete Bewegung zum Ausdruck bringen, wird die Richtungsangabe häufig nicht als freie Fügung, sondern als Valenz angesehen. Beispiele aus HaGenLex sind *laufen* und *fahren*. Positiv gewendet scheint sich damit ein Hilfsmittel zur Aufdeckung direktionaler Valenzen anzudeuten. Hier muss aber sofort einschränkend darauf hingewiesen werden, dass Verbpartikeln wie *herum* (oder auch *vorbei* und *entlang*) direktionale Angaben blockieren:[9] **Peter fuhr in die Stadt herum*. Andererseits ist aber *herumfahren.1* Hyponym von *fahren.4*. Hier wäre eine sorgfältige Kreuzklassifikation bezüglich der durch Partikeln ausgedrückten Art und Weise der Bewegung hilfreich.[10]

Abschließend sei exemplarisch auf ein nicht untypisches Einzelproblem hingewiesen. Der Versuch, diejenigen Verben einzugrenzen, die direktionale Ergänzungen als freie Fügungen erlauben, führt unter anderem zum Ausschlusskonzept *?bewegen_auf_Stelle.1*. Hierunter findet man *?iterative_Bew.2* und darunter *tanzen.2*, das als einzig mögliche Lesart für *Das Paar tanzte ins Nachbarzimmer* in Frage kommt. In derartigen Fällen ist es nicht immer unmittelbar einsichtig, ob nur eine Fehlklassifikation in GermaNet vorliegt, oder ob es sich um ein tiefergehendes Problem in der semantischen Hierarchie handelt.

4 Zusammenfassung

Am Beispiel des semantikbasierten Lexikons HaGenLex wurde illustriert, wie eine Lesarten-Kopplung mit GermaNet die Pflege und Erweiterung anderer lexikalischer Ressourcen unterstützen kann, und wie sich dabei gleichzeitig Schwachstellen von GermaNet aufdecken lassen.

Insbesondere hat sich gezeigt, dass eine Restrukturierung resp. Ergänzung der GermaNet-Hierarchie hinsichtlich Aktionsart und Aspekt aus Sicht von HaGenLex sehr wünschenswert wäre. Das Gleiche lässt sich von einer Kreuzklassifikation von Partikelverben hinsichtlich des semantischen Beitrags der Partikeln sagen.

Anhang: Interne Repräsentation

Die interne Repräsentation von HaGenLex-Einträgen basiert auf einem Merkmal-Wert-Formalismus im Stile von (Carpenter 1992). Das Kerngerüst bildet eine baumförmige Typhierarchie, wobei zu jedem Typ angegeben ist, welche Merkmale mit welchen Werten für Strukturen dieses Typs zulässig sind. In Abbildung 1 sind einige der wichtigsten in HaGenLex verwendeten Merkmalsdeklarationen (in leicht vereinfachter Form) aufgeführt.

Alle lexikalischen Einträge sind vom Typ *word*. Da *word* in der Typhierarchie unter *sign* angeordnet ist, „erbt" jede Struktur vom Typ *word* per Konvention die für *sign* deklarierten

Merkmale. Die oberste Merkmalsebene eines Eintrags ergibt sich somit aus der Zusammenfassung der für *word* und *sign* deklarierten Merkmale; vgl. Abbildung 2. Neben Angaben zu Morphologie (MORPH) und Syntax (SYN) findet sich hier das Merkmal SEMSEL, das eine Struktur vom Typ *semsel* zum Wert hat, die ihrerseits die Semantik (SEM) und die Valenz (SELECT) des Eintrags näher spezifiziert. Die Valenzinformation wiederum besteht aus einer Liste von Strukturen des Typs *select-element*. Jede dieser Strukturen bringt die semantische Rolle des jeweiligen Arguments zum Ausdruck (REL), dessen syntaktische Notwendigkeit (OBLIG) sowie seine weitere Charakterisierung als Struktur vom Typ *sign*. Strukturen vom Typ *sem* schließlich kodieren die Semantik eines Eintrags durch seine semantische Sorte (ENTITY), zusätzliche MultiNet-Spezifikationen (NET), MultiNet-Merkmale des Konzeptknotens (LAY) sowie durch den Hinweis, ob eine bestimmte Ausprägung regulärer Polysemie vorliegt (MOLEC).

sign

MORPH	*morph*
SYN	*syn*
SEMSEL	*semsel*

word

G-ID	*set*(*integer*)
ORIGIN	*string*

semsel

SEM	*sem*
C-ID	*string*
DOMAIN	*domain*
SELECT	*list*(*select-element*)
COMPAT-R	*set*(*rel*)

sem

ENTITY	*entity*
NET	*net*
LAY	*lay*
MOLEC	*boolean*

select-element

REL	*set*(*rel*)
OBLIG	*boolean*
SEL	*sign*

Abbildung 1: Ausschnitt der in HaGenLex verwendeten Merkmalsdeklarationen.

Gegenwärtig sind Merkmal-Wert-Strukturen im Rahmen von HaGenLex als Scheme-Strukturen implementiert, auf die sowohl die Lexikonwerkzeuge als auch der Parser über gemeinsame Schnittstellen zugreifen.

Anmerkungen

1 Für eine detaillierte Darstellung sei der Leser auf HELBIG 2001 verwiesen.

2 Man beachte, dass hier nicht nur generische Konzepte gemeint sind, sondern dass etwa auch *der diesjährige GermaNet-Workshop in Tübingen* als ein Konzept aufgefasst wird, das einem Knoten in der zugehörigen semantischen Repräsentation entspricht.

3 Eine Beschreibung des dabei verwendeten Parsers gibt HARTRUMPF 2003, Kap. 3. Eine Anwendung zur natürlichsprachlichen Informationsrecherche wird in LEVELING & HELBIG 2002 vorgestellt.

4 Alle Verweise beziehen sich auf GermaNet 4.0.

5 Vgl. etwa KUNZE & WAGNER 2001.

6 Nach anfänglichen Versuchen einer vollautomatischen Lesartenzuordnung hat sich letztlich die Software-unterstützte Zuordnung durch den Lexikographen als effektivste und verlässlichste Methode erwiesen.

7 Als Nebenprodukt einer solchen, an übergeordneten semantischen Gesichtspunkten orientierten Sichtung der Struktur von GermaNet ergeben sich wiederum Hinweise auf Schwachstellen derselben. Problematisch erscheint es beispielsweise, *aufwachen.1* und *einschlafen.3*, neben *schlummern.1* und *dösen.1*, als Hyponyme von *schlafen.2* anzusehen. Ferner sind *platzen.1* und *zerschellen.1* nicht nur Hyponyme von *Mat_Zustands_Veränderung.1*, sondern auch von {*zerstören.1, destruieren.1*}, was wiederum *Mat_Zustands_Veränderung.2* untergeordnet ist. Andererseits soll aber *Mat_Zustands_Veränderung.2*, im Gegensatz zu *Mat_Zustands_Veränderung.1*, gerade *kausative* Veränderungen der materiellen Beschaffenheit erfassen.

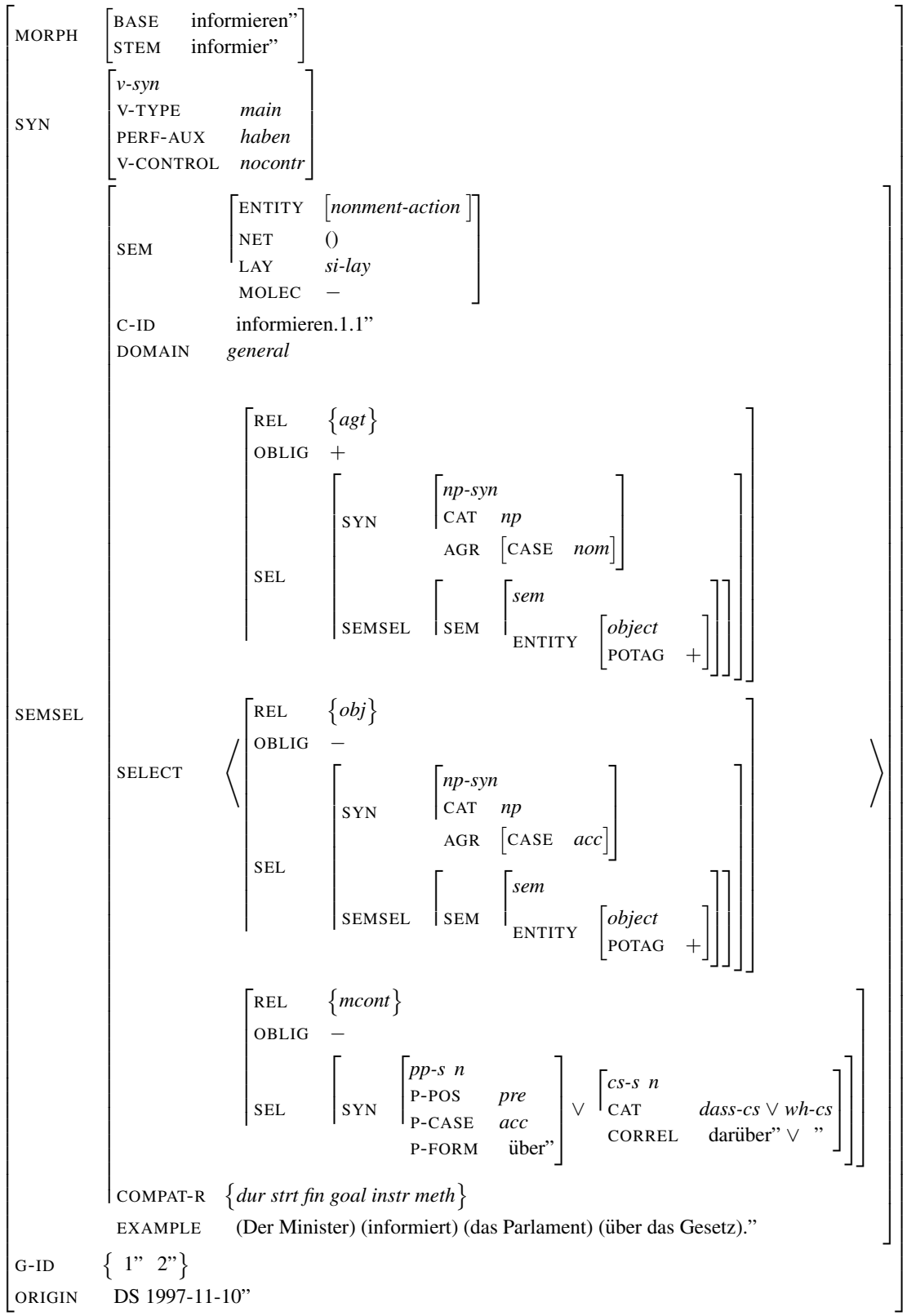

Abbildung 2: Merkmal-Wert-Struktur des HaGenLex-Eintrags für *informieren*.

[8] Ähnlich verhält es sich übrigens mit bestimmten Kognitionsverben wie *blicken*.

[9] Die unter *http: //www.sfs.nphil.uni-tuebingen.de/lsd/* zu findende Online-Dokumentation zu GermaNet führt eine derartige Klassifikation im Abschnitt "Future Work" auf – ebenso wie die Klassifikation nach Aktionsarten [Zugriff April 2004].

Literatur

Carpenter, B. (1992). The Logic of Typed Feature Structures. Cambridge Tracts in Theoretical Computer Science 32. Cambridge University Press.

Fellbaum, Ch. (ed.) (1998). WordNet – An Electronic Lexical Database. Language, Speech, and Communication. Cambridge, MA / London: MIT Press.

Hartrumpf, S.; Helbig, H.; Osswald, R. (2003). "The Semantically Based Computer Lexicon HaGenLex – Structure and Technological Environment." In: Traitement Automatique des Langues 44(2) (2003) [erscheint].

Hartrumpf, S. (2003). Hybrid Disambiguation in Natural Language Analysis. Osnabrück: Der Andere Verlag.

Helbig, H. (2001). Die semantische Struktur natürlicher Sprache: Wissensrepräsentation mit MultiNet. Berlin et al.: Springer.

Kunze, C.; Wagner, A. (2001). „Anwendungsperspektiven des GermaNet, eines lexikalisch-semantischen Netzes für das Deutsche." In: Lemberg, I.; Schröder, B.; Storrer, A. (Hrsg.) (2001). Chancen und Perspektiven computergestützter Lexikographie. Tübingen: Niemeyer [= Lexicographica Series Maior Vol. 107], 229-246.

Leveling, J.; Helbig, H. (2002). "A Robust Natural Language Interface for Access to Bibliographic Databases." In: Callaos, N.; Margenstern, M.; Sanchez, B. (eds.) (2002). Proceedings of the 6th World Multiconference on Systemics, Cybernetics and Informatics (SCI 2002), volume XI. Orlando, FL: International Institute of Informatics and Systemics (IIIS), 133-138.