

1 Draft

1.1 Motivation von Merkmalstrukturen

1.1.1 Merkmalstrukturen ermöglichen Generalisierungen

Ein Kennzeichen aller natürlicher Sprachen ist, daß sie eine unendliche Zahl von Sätzen umfassen. Eine linguistische Theorie, die eine natürliche Sprache beschreiben soll, muß demnach Produktionsregeln bereithalten, mit deren Hilfe unendlich viele wohlgeformte Ausdrücke gebildet werden können. Dies ist bei einer endlichen Zahl von Regeln nur möglich, wenn den linguistischen Objekten, z.B. den Wörtern, Klassen zugeordnet werden, über die die Regeln generalisieren.¹

Hierbei ist die Verwendung von Merkmalen zur Klassifizierung linguistischer Objekte einer arbiträren Klassenzuweisung in vieler Hinsicht überlegen. Sie ermöglicht es zum einen die Unterschiede zwischen zwei Objekten verschiedener Klassen explizit zu beschreiben. Zum anderen wird durch die Merkmale eine partielle Ordnung auf der Menge der Klassen definiert, die das Konzept von Teilklasse und allgemeinerer Klasse modelliert. Darüberhinaus kann jede Klasse, die wenigstens zwei verschieden Objekte umfaßt, durch die Hinzunahme weiterer Merkmale in Teilklassen zerlegt werden, d.h. die Domäne kann so explizit beschrieben werden, wie gewünscht. Klassische merkmalsbasierte Grammatiktheorien verwenden ausschließlich binäre Merkmale, wie z.B. [\pm singular].² Ihren Ausgangspunkt findet die Verwendung von Merkmalen zur Beschreibung linguistischer Objekte in der strukturalistischen Phonologie (siehe Bloomfield, 1933; Trubetzkoy, 1939), die Phoneme als Bündel distinktiver Merkmale auffaßt (siehe hierzu Jakobson and Halle, 1956); diese Auffassung findet sich auch in moderneren Ansätze im Rahmen der generativen Phonologie (z.B. Kenstowicz and Kisseberth, 1979), die durch die Arbeiten von Chomsky und Halle zur Phonologie des Englischen begründet wurde (siehe Chomsky and Halle, 1968). Gibt man die Beschränkung auf binäre Merkmale auf und versteht Merkmale als Attribut-Wert-Paare, so ergeben sich zusätzliche Möglichkeiten Redundanzen in einer Grammatik zu vermeiden.

Im folgenden soll an einem Beispiel verdeutlicht werden, wie mit Hilfe von Attribut-Wert-Paaren Generalisierungen ausgedrückt werden können. Wir betrachten dazu den folgenden Ausschnitt aus einer kontextfreien Grammatik des Spanischen, die auf die Verwendung von Attribut-Wert-Paaren verzichtet und ausschließlich mit arbiträren Klassenbezeichnern, bzw. Nichtterminalsymbolen arbeitet:

¹“Clearly the set of compound linguistic expressions in a natural language is not finite, so we cannot list them. An interpreted formal system defining the membership of the collection of linguistic expressions, and assigning a structure and an interpretation to each member, is required. This is what we call a grammar.”(Gazdar et al., 1985, Seite 1).

²Siehe unter anderem Chomsky (1965).

S → NP1sg VP1sg	<i>yo</i> ↦ NP1sg	<i>hablo</i> ↦ VP1sg
S → NP2sg VP2sg	<i>tú</i> ↦ NP2sg	<i>hablas</i> ↦ VP2sg
S → NP3sg VP3sg	<i>ella</i> ↦ NP3sg	<i>habla</i> ↦ VP3sg
S → NP1pl VP1pl	<i>nosotros</i> ↦ NP1pl	<i>hablamos</i> ↦ VP1pl
S → NP2pl VP2pl	<i>vosotros</i> ↦ NP2pl	<i>habláis</i> ↦ VP2pl
S → NP3pl VP3pl	<i>ellas</i> ↦ NP3pl	<i>hablan</i> ↦ VP3pl

Diese Grammatik benötigt sechs Regeln, um die Struktur der folgenden sechs Sätze zu erfassen:

yo hablo
tú hablas
ella habla
nosotros hablamos
vosotros habláis
ellas hablan

Auch wenn das Zahlenverhältnis von 6 Regeln zu 6 erfaßten Sätzen etwas anderes suggeriert, so kommen in dieser Beispielgrammatik durchaus Generalisierungen zum Ausdruck. Erweitert man das Lexikon um die Formen des Worts ‘cantar’, so erlaubt diese Grammatik zum Beispiel auch die korrekte Analyse des Satzes ‘yo canto’. Die Grammatik muß jedoch an der Aufgabe, die Subjekt-Verb-Kongruenz systematisch zu beschreiben, scheitern. So tragen die beiden Nichtterminalsymbole NP1sg und VP1sg zwar scheinbar beide die Information ‘erste Person’ und ‘singular’, formal gesehen, handelt es sich jedoch um zwei völlig unabhängige atomare Symbole, die nur aus mnemotechnischen Gründen eine ähnliche Bezeichnung tragen. Es ist also innerhalb der vorgestellten Grammatik nicht möglich über das Kongruenzverhalten der Klassen NP1sg, NP2sg, NP3sg, NP1pl, ... und VP1sg, VP2sg, VP3sg, VP1pl, ... zu generalisieren.

Erst durch den Einsatz von Attribut-Wert-Paaren wird es möglich, über die sechs Regeln zu generalisieren, die zusammen die Subjekt-Verb-Kongruenz in einem spanischen Satz beschreiben. Anstatt ein Merkmal [\pm plural] anzunehmen, wird ein Attribut NUM eingeführt, das die Werte **sing** und **plur** annehmen kann. Die Subjekt-Verb-Kongruenz kann nun in Form zweier Constraints beschrieben werden, die garantieren, daß die Verbal- und die Nominalphrase eines Satzes sowohl dem Attribut NUM, als auch dem Attribut PERS den gleichen Wert zuordnen. Formuliert man diese Constraints in der im Grammatikformalismus PATR-II³ verwendeten Syntax, so erhält man folgende Grammatikregel:⁴

³PATR-II (*parsing and translation*) ist ein von Shieber entwickelter Grammatikformalismus, für die Formulierung von Unifikationsgrammatiken (siehe Shieber, 1983).

⁴Die PATR-Regel zerfällt in zwei Teile: eine gewöhnliche kontextfreie Phrasenstrukturregel, die die lineare Abfolge und die Kategorie der Konstituenten festlegt und einer Menge von Constraints, angegeben in Form von Gleichungen, die die Gleichheit der Werte zweier Attribute fordern.

$$\begin{aligned}
S &\rightarrow NP VP \\
\langle VP \text{ NUM} \rangle &= \langle NP \text{ NUM} \rangle \\
\langle VP \text{ PERS} \rangle &= \langle NP \text{ PERS} \rangle
\end{aligned}$$

Diese generalisierte Satzregel wird durch komplexere lexikalische Einträge ermöglicht, die jedem Wort statt eines atomaren Kategoriebezeichners eine Merkmalstruktur zuordnen. Diese Verschiebung der linguistischen Information ins Lexikon ist ein Trend, der in den meisten modernen linguistischen Theorien zu beobachten ist.⁵ Das Lexikon zu unserer Beispielgrammatik wird somit zu:

$$\begin{array}{ccc}
yo & \mapsto & \begin{bmatrix} \text{CAT:} & \text{NP} \\ \text{NUM:} & \text{sing} \\ \text{PERS:} & 1 \end{bmatrix} & hablo & \mapsto & \begin{bmatrix} \text{CAT:} & \text{VP} \\ \text{NUM:} & \text{sing} \\ \text{PERS:} & 1 \end{bmatrix} \\
tú & \mapsto & \begin{bmatrix} \text{CAT:} & \text{NP} \\ \text{NUM:} & \text{sing} \\ \text{PERS:} & 2 \end{bmatrix} & hablas & \mapsto & \begin{bmatrix} \text{CAT:} & \text{VP} \\ \text{NUM:} & \text{sing} \\ \text{PERS:} & 2 \end{bmatrix} \\
ella & \mapsto & \begin{bmatrix} \text{CAT:} & \text{NP} \\ \text{NUM:} & \text{sing} \\ \text{PERS:} & 3 \end{bmatrix} & habla & \mapsto & \begin{bmatrix} \text{CAT:} & \text{VP} \\ \text{NUM:} & \text{sing} \\ \text{PERS:} & 3 \end{bmatrix} \\
& \vdots & & & \vdots &
\end{array}$$

Die Grammatik kann weiter vereinfacht werden, indem man zuläßt, daß die Attributwerte in Merkmalstrukturen selber wieder komplexe Merkmalstrukturen sein können. Dadurch ist es möglich, ein Kongruenzattribut AGR (*agreement*) einzuführen, dessen Werte Merkmalstrukturen sind, die sich aus den kongruierenden Attribut-Wert-Paaren aufbauen.

$$\begin{aligned}
S &\rightarrow NP VP \\
\langle VP \text{ AGR} \rangle &= \langle NP \text{ AGR} \rangle
\end{aligned}$$

⁵“Lexical entries have evolved from simple pairings of phonological forms with grammatical categories into elaborate information structures, in which phonological forms are now paired with more articulated feature structure descriptions.”(Sag and Wasow, 1999, Seite 173).

$yo \mapsto$	$\left[\begin{array}{l} \text{CAT: NP} \\ \text{AGR: } \left[\begin{array}{l} \text{NUM: sing} \\ \text{PERS: 1} \end{array} \right] \end{array} \right]$	$hablo \mapsto$	$\left[\begin{array}{l} \text{CAT: VP} \\ \text{AGR: } \left[\begin{array}{l} \text{NUM: sing} \\ \text{PERS: 1} \end{array} \right] \end{array} \right]$
$tú \mapsto$	$\left[\begin{array}{l} \text{CAT: NP} \\ \text{AGR: } \left[\begin{array}{l} \text{NUM: sing} \\ \text{PERS: 2} \end{array} \right] \end{array} \right]$	$hablas \mapsto$	$\left[\begin{array}{l} \text{CAT: VP} \\ \text{AGR: } \left[\begin{array}{l} \text{NUM: sing} \\ \text{PERS: 2} \end{array} \right] \end{array} \right]$
$ella \mapsto$	$\left[\begin{array}{l} \text{CAT: NP} \\ \text{AGR: } \left[\begin{array}{l} \text{NUM: sing} \\ \text{PERS: 3} \end{array} \right] \end{array} \right]$	$habla \mapsto$	$\left[\begin{array}{l} \text{CAT: VP} \\ \text{AGR: } \left[\begin{array}{l} \text{NUM: sing} \\ \text{PERS: 3} \end{array} \right] \end{array} \right]$
\vdots	\vdots	\vdots	\vdots

1.1.2 Aufbau und Darstellung von Merkmalstrukturen

Merkmalstrukturen kommen in unterschiedlichen Disziplinen unter verschiedenen Bezeichnungen zum Einsatz. Im Rahmen der Künstlichen Intelligenz und der kognitiven Psychologie (siehe Barsalou, 1992) dienen sie in Form von *Frames* der Repräsentation konzeptuellem Wissens. Auch bei den in Programmiersprachen häufig zum Einsatz kommenden *record structures* handelt es sich um Merkmalstrukturen. Die verschiedenen merkmalsstruktur-basierten Formalismen stellen unterschiedliche Anforderungen an die Strukturen. Fundamental gemeinsam ist jedoch allen Ansätzen, daß Merkmalstrukturen als endliche partielle Funktion von Attributen nach Attributwerten aufgefaßt werden. Prinzipiell lassen sich Merkmalstrukturen in zwei Formen darstellen, als Attribut-Wert-Matrix (*attribute value matrix, AVM*) wie in unserem Beispielllexikon oder als beschrifteter gerichteter Graph,⁶ siehe Abbildung 1.

An den Graphen können unterschiedliche Anforderungen gestellt werden, entweder es werden nur *Bäume* zugelassen wie in Abbildung 1 oder beliebige gerichtete beschriftete Graphen wie in Abbildung 2⁷. Die Graphen (b) und (c) in Abbildung 2 sind keine Bäume, da sie Knoten mit mehr als einer eingehenden Kante haben. In den Strukturen (a) und (b) werden den Attributen die gleiche Werte zugeordnet, jedoch haben die Attribute D und E in der Struktur (b) nicht nur den gleichen Wert, sondern die mit diesen Attributen beschrifteten Kanten verweisen auf die gleiche Ecke, d.h. die Werte sind nicht nur *typidentisch* wie in Abbildung 2(a) sondern *tokenidentisch* und haben somit denselben Referenten. Um die nicht baumförmigen Merkmalstrukturen 2(b) und 2(c) als Attribut-Wert-

⁶Eine Einführung in die Graphentheorie findet sich im Anhang ??.

⁷Da Merkmalstrukturen als Funktionen von Attributen nach Werten aufgefaßt werden, sind die Graphen nicht wirklich beliebig: Von keiner Ecke dürfen zwei Kanten mit gleicher Beschriftung ausgehen (siehe Abbildung 3).

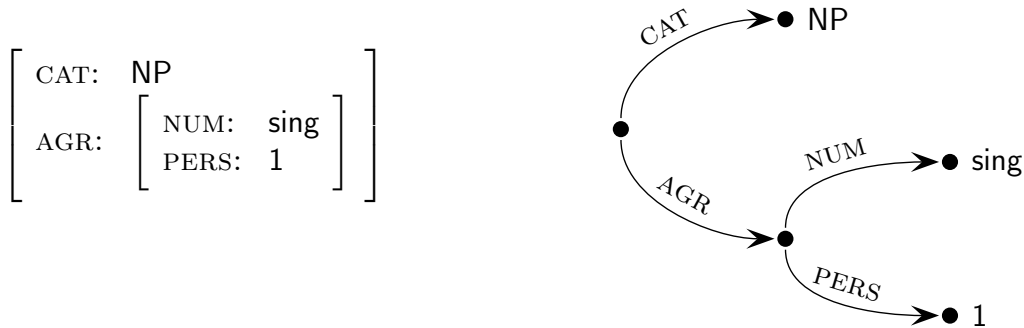


Figure 1: Merkmalstruktur zu *yo*, dargestellt als AVM und als beschrifteter gerichteter Graph.

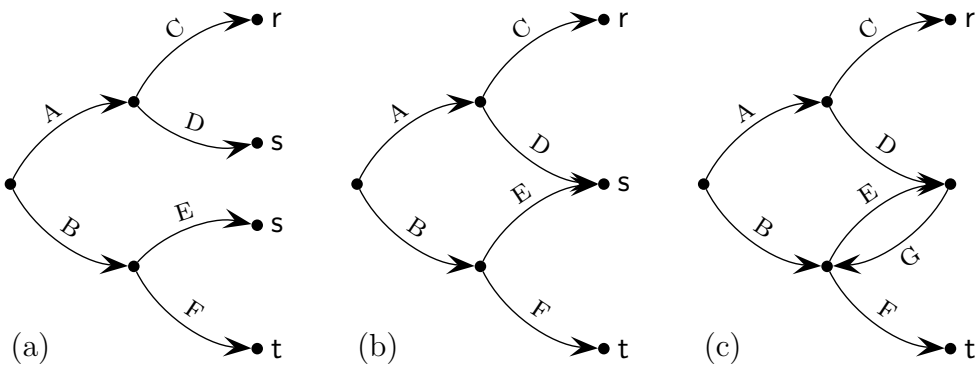


Figure 2: Gerichtete beschriftete Graphen die Merkmalstrukturen repräsentieren.

Matrizen darzustellen, verwenden wir wie in Abbildung 4 Indizes: Koindizierungen, d.h. identische Indizes, markieren dabei Tokenidentität bzw. Koreferenz.

Der Graph 2(c) enthält einen Zyklus, d.h. es gibt eine Ecke, von der aus es möglich ist, entlang gerichteter Kanten einen Weg zu beschreiten, der zum Ausgangspunkt zurückführt. Merkmalstrukturen, die durch *zyklische* Graphen repräsentiert werden, nennt man *zyklische Merkmalstrukturen*. Zahlreiche Formalismen schließen solche Strukturen aus.

1.1.3 Subsumption, Unifikation und Generalisierung

Auf der Menge der Merkmalstrukturen läßt sich eine natürliche partielle Ordnung,⁸ die die Informativität der Strukturen widerspiegelt, definieren; diese Ordnung wird *Subsumptionsrelation* genannt. Wir sagen, daß eine Merkmalstruktur \mathcal{A} eine Merkmalstruktur \mathcal{B} subsumiert, wenn \mathcal{B} informativer, bzw. spezifischer ist als \mathcal{A} ; wir schreiben dann $\mathcal{A} \sqsubseteq \mathcal{B}$. Abbildung 5 zeigt zwei Merkmalstrukturen, von denen die eine die andere subsumiert.

⁸Der Begriff der partiellen Ordnung wird im Anhang ?? definiert.

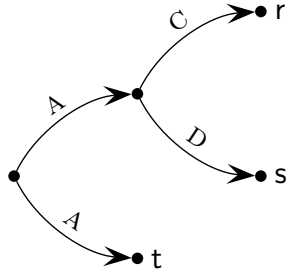


Figure 3: Gerichteter beschrifteter Graph der keine Merkmalstruktur repräsentiert.

$$\begin{array}{ccc}
 \left[\begin{array}{l} \text{A:} \\ \text{B:} \end{array} \left[\begin{array}{l} \text{C: } r \\ \text{D: } s \end{array} \right] \right] & \left[\begin{array}{l} \text{A:} \\ \text{B:} \end{array} \left[\begin{array}{l} \text{C: } r \\ \text{D: } \boxed{1}s \end{array} \right] \right] & \left[\begin{array}{l} \text{A:} \\ \text{B:} \end{array} \left[\begin{array}{l} \text{C: } r \\ \text{D: } \boxed{1} \left[\text{G: } \boxed{2} \right] \end{array} \right] \right] \\
 \text{(a)} & \text{(b)} & \text{(c)}
 \end{array}$$

Figure 4: Merkmalstrukturen aus Abbildung 2 dargestellt als AVMs.

$$\left[\begin{array}{l} \text{A:} \\ \text{B:} \end{array} \left[\begin{array}{l} \text{C: } r \\ \text{D: } s \end{array} \right] \right] \sqsubseteq \left[\begin{array}{l} \text{A:} \\ \text{B:} \end{array} \left[\begin{array}{l} \text{C: } r \\ \text{D: } s \end{array} \right] \right] \sqsubseteq \left[\begin{array}{l} \text{A:} \\ \text{B:} \end{array} \left[\begin{array}{l} \text{C: } r \\ \text{D: } \boxed{1}s \end{array} \right] \right]$$

Figure 5: Beispiel für die Subsumptionsrelation

$$\left[\begin{array}{l} \text{A: } \boxed{1} \\ \text{B: } \boxed{1} \end{array} \left[\begin{array}{l} \text{C: } r \\ \text{D: } s \end{array} \right] \right] \sqcup \left[\begin{array}{l} \text{A: } \left[\begin{array}{l} \text{C: } r \end{array} \right] \\ \text{B: } \left[\begin{array}{l} \text{E: } r \end{array} \right] \\ \text{F: } t \end{array} \right] = \left[\begin{array}{l} \text{A: } \boxed{1} \\ \text{B: } \boxed{1} \\ \text{F: } t \end{array} \left[\begin{array}{l} \text{C: } r \\ \text{D: } s \\ \text{E: } r \end{array} \right] \right]$$

Figure 6: Beispiel einer Unifikation.

$$\left[\begin{array}{l} \text{A: } \boxed{1} \\ \text{B: } \boxed{1} \end{array} \left[\begin{array}{l} \text{C: } r \\ \text{D: } s \end{array} \right] \right] \sqcap \left[\begin{array}{l} \text{A: } \left[\begin{array}{l} \text{C: } r \end{array} \right] \\ \text{B: } \left[\begin{array}{l} \text{E: } r \end{array} \right] \\ \text{F: } t \end{array} \right] = \left[\text{A: } \left[\begin{array}{l} \text{C: } r \end{array} \right] \right]$$

Figure 7: Beispiel einer Generalisierung.

Zu zwei Merkmalstrukturen mit kompatibler Information kann eine dritte Struktur gebildet werden, die die Information beider Strukturen miteinander kombiniert. Die Operation der Kombination nennt man *Unifikation*, dabei ist das Resultat der Unifikation die Merkmalstruktur, die man erhält, wenn man die Bestandteile beider Ausgangsstrukturen zusammenfügt; die Unifikation zweier Merkmalstrukturen ist somit die bezüglich der Subsumptionsrelation allgemeinste Merkmalstruktur, die von beiden Ausgangsstrukturen subsumiert wird. Abbildung 6 zeigt ein Beispiel einer Unifikation zweier Merkmalstrukturen. Der Versuch zwei Merkmalstrukturen mit inkompatibler Information zu unifizieren scheitert.

Die der Unifikation entgegengesetzte Operation ist die *Generalisierungsoperation*. Sie nimmt zwei Merkmalstrukturen und liefert eine neue Struktur, in der alle Informationen enthalten sind, die von beiden Ausgangsstrukturen geteilt werden. Somit ist das Ergebnis der Generalisierung zweier Merkmalstrukturen die spezifischste Struktur, die beide Ausgangsstrukturen subsumiert. Diese Operation gelingt immer, wenn davon ausgegangen wird, daß es eine allgemeinste Merkmalstruktur $\left[\quad \right]$ gibt. Abbildung 7 zeigt ein Beispiel für die Generalisierungsoperation.

Zusammenfassend läßt sich sagen, daß Merkmalstrukturen Datenstrukturen sind, die komplexe Information erfassen können. Dies gelingt, da rekursive Einbettungen möglich sind und Strukturen geteilt werden können. Merkmalstrukturen können bezüglich ihrer Informativität partiell geordnet werden.

1.1.4 Was repräsentieren Merkmalstrukturen?

Im Rahmen von linguistischen Theorien finden sich zwei prinzipiell verschiedene Verwendungen von Merkmalstrukturen: Sie werden zum einen eingesetzt, um linguistische Objekte zu repräsentieren oder modellieren; zum anderen dienen sie der Repräsentation von Beschreibungen linguistischer Objekte. Im ersten Fall, wenn eine Merkmalstruktur ein linguistisches Objekt direkt repräsentiert, werden die Strukturen in der Regel als komplett bzw. vollständig instantiiert aufgefaßt. Im Rahmen von unifikationsbasierten Grammatikformalismen, die speziell für die maschinelle Sprachverarbeitung entwickelt wurden, wird hingegen häufig die Sichtweise eingenommen (siehe Shieber, 1986; Kasper and Rounds, 1990, ???), daß nicht ein linguistisches Objekt selbst, sondern die Information über das Objekt, die zu einem bestimmten Zeitpunkt der Sprachverarbeitung vorhanden ist, durch eine Merkmalstruktur repräsentiert wird.⁹ Diese Auffassung wird auch in der frühen HPSG (Pollard and Sag, 1987) vertreten:

“Intuitively, a feature structure is just an information-bearing object that describes or represents another thing by specifying *values* for various *attributes* of the described thing; we think of the feature structure as providing partial information about the thing described.” (siehe Pollard and Sag, 1987, Seite 28)

In der späteren HPSG (Pollard and Sag, 1994; Sag and Wasow, 1999) hingegen wird zwischen Merkmalstrukturen, die linguistische Objekte vollständig spezifizieren, solchen die die vorhandene Information über Objekte repräsentieren und Beschreibungen von Merkmalstrukturen unterschieden:

“Feature structures that serve as models of linguistic entities are required to satisfy further criteria of completeness. Roughly what this means is that they are total (not merely partial) models of the objects that they represent.” (siehe Pollard and Sag, 1994, Seite 18)

Und weiter zur Repräsentation der Beschreibungen von Merkmalstrukturen:

“A common source of confusion is that feature structures themselves can be used as descriptions of other feature structures. [...] We choose to eliminate this possible source of confusion by using only totally well-typed, sort-resolved feature structures as (total) models of linguistic entities and AVM diagrams (not feature structures) as descriptions.” (siehe Pollard and Sag, 1994, Seite 21)

⁹“From a unification-based or ‘information-processing’ perspective, a feature structure may be viewed as representing just that information which is available about a linguistic object at a particular point in a computation.[...] Alternatively, it is possible to view feature structures as complete or ‘totally-instantiated’ representations of linguistic objects.”(Keller, 1993, Seite 12).

Da die Unifikationsoperation nur unter der Annahme partiell verstandener Merkmalstrukturen Sinn macht, finden sich in allen unifikationsbasierten Formalismen Merkmalstrukturen (im Falle der späteren HPSG heißen sie AVMs, *attribute value matrices*), die zur partiellen Beschreibung bestimmter linguistischer Objekte eingesetzt werden. So werden z.B. in der LFG (siehe Bresnan 1982 und darin insbesondere Kaplan and Bresnan 1982) grammatikalische Funktionen in Form von partiell aufgefaßten Merkmalstrukturen beschrieben, während die linguistischen Zeichen als vollständig instantiierte Strukturen analysiert werden.

Moderne linguistische Theorien gehen von der Grundannahme aus, daß die Grammatik einer Sprache ein *deklaratives* System von Constraints sein sollte. Somit sollte eine Grammatik unabhängig von der prozeduralen Strategie der Sprachverarbeitung sein. Das hat zur Folge, daß die Wahl einer anderen Parsing-technik nicht dazu führen kann, daß eine andere Menge von Sätzen als wohlgeformt analysiert wird. Um dies zu erreichen, darf die Grammatik nur Operationen umfassen, die zu einem monotonen Informationszuwachs führen. In unifikationsbasierten Theorien wird daher von der Annahme ausgegangen, daß in jedem Sprachverarbeitungsprozeß (Sprachproduktion, Sprachperzeption, ...) sich die exakte Spezifikation eines linguistischen Tokens aus einem kumulativen Prozeß ergibt, in dem verschiedene Constraints interagieren (lexikalische Einträge, grammatische Regeln, kontextuelle Faktoren); die Spezifikation ergibt sich durch Unifikation der Information aus den verschiedenen Quellen. Somit ist die Monotoniebedingung notwendig für unifikationsbasierte Formalismen.

Pollard and Sag (1987, Seite 8) erklären hierzu, daß deklarative, partielle Information die in beliebiger Reihenfolge angereichert werden kann, ohne das Ergebnis zu beeinflussen, in einem Sprachverarbeitungsprozeß flexibel aufgerufen werden kann. Da der Informationszuwachs monoton ist und somit Operationen, die Information zerstören unzulässig sind, bleibt die partielle Information, die zu einem linguistischen Token bekannt ist, wahr, egal wieviel zusätzliche Information hinzukommt. Ein derartiges System macht die Information parallel zugänglich für allen möglichen Prozessen der Sprachverarbeitung.

Zusammenfassend läßt sich festhalten, daß wenn eine Merkmalstruktur eine Beschreibung eines linguistischen Objekts oder die Information, die zu einem bestimmten Zeitpunkt der Sprachverarbeitung über ein linguistisches Objekt vorhanden ist, repräsentiert, so wird sie notwendigerweise als inhärent partiell verstanden, das heißt, es wird davon ausgegangen, daß zu einem späteren Zeitpunkt zusätzliche Information hinzukommen kann. So können zwei partiell verstandene Merkmalstrukturen verschiedene Aspekte des selben linguistischen Objekts repräsentieren, die Verschmelzung der Information beider Strukturen erfolgt mittels Unifikation. Da in allen unifikationsbasierten Ansätzen Merkmalstrukturen notwendigerweise immer als partielle Repräsentationen von linguistischen Objekten aufgefaßt werden müssen, werden wir diese Perspektive auch im Rahmen dieser Arbeit einnehmen.

1.1.5 Das Problem der Unterspezifikation

Eine Merkmalstruktur, die nur bestimmte Aspekte eines Objekts repräsentiert, ist prinzipiell unterspezifiziert, weitere Attribut-Wert-Paare können auf das Objekt zutreffen, ohne in der Merkmalstruktur zu erscheinen. Gerade aus der Sicht der Sprachverarbeitung bietet die Annahme, daß Merkmalstrukturen unterspezifiziert sein können, Vorteile. Trifft man in einem Sprachverarbeitungsprozeß auf den Ausdruck ‘Jungen’, so kann mit Hilfe des Lexikons nur die Informa-

tion $\left[\begin{array}{l} \text{CAT : N} \\ \text{GEN : masc} \end{array} \right]$ über die Kategorie und den Genus ermittelt werden, die

Kasus- und Numerusinformation ergibt sich erst aus dem jeweiligen Kontext, z.B. ‘Paula sieht den Jungen’, und muß mittels Unifikation in die Gesamtrepräsentation eingebracht werden. Zusätzlich erlaubt erst die Unterspezifikation die Bildung von allgemeineren Klassen linguistischer Objekte, wie z.B. $\left[\text{CAT : N} \right]$, der Klasse der Nomen. Diese Fähigkeit solcher Formalismen zur flexiblen Klassenbildung ermöglicht kompaktere Repräsentationen, da über Klassen generalisiert werden kann.

Allerdings müssen Formalismen, die Unterspezifikation zulassen, mit zusätzlichen Schwierigkeiten kämpfen: Ein Problem partiell verstandener Merkmalstrukturen ist, daß keine Unterscheidung zwischen nichtspezifizierten Attributen und irrelevanten Attributen möglich ist. Taucht in einer partiell verstandenen Merkmalstruktur ein Attribut nicht auf, so kann daraus nicht geschlossen werden, daß für das beschriebene Objekt dieses Attribut irrelevant oder sogar inadäquat ist; dies führt insbesondere zu Problemen, wenn negative Constraints über Attribute und ihre Werte ausgedrückt werden sollen. Abschnitt 1.2 wird sich ausführlicher mit diesem Problem auseinandersetzen.

Zum Abschluß dieses Abschnitts soll mit dem folgenden Zitat noch einmal die besondere Leistungsfähigkeit von Merkmalstrukturen, im Gegensatz zum Beispiel zu prädikatenlogischen Ausdrücken, herausgestellt werden, die darauf beruht, daß gezielt auf Substrukturen verwiesen und über diese generalisiert werden kann:

“The principal benefit of feature structures is that they provide named access to properties or substructures in the formal representation of an entity by means of these paths. This is in contrast to first-order terms, for example, whose subterms are referred to by means of ordinals: first argument, second argument of the first argument, etc.” (siehe Penn, 2000, Seite 2)

1.2 Motivation von getypten Merkmalstrukturen und Typsignaturen

1.2.1 Vorbemerkung

Bereits im vorangegangenen Abschnitt haben wir gesehen, daß partiell aufgefaßte Merkmalstrukturen keine Unterscheidung zwischen irrelevanten und unterspezifizierten Attributen ermöglichen. Dies führt zu Problemen bei der Entwicklung von Grammatiken.

Die Aufgabe einer Grammatik ist es die Klasse der wohlgeformten bzw. grammatikalischen Ausdrücke in der Klasse aller möglichen Ausdrücke zu identifizieren. Die Beschränkung auf die zulässigen Merkmalstrukturen bedarf in unifikationsbasierten Formalismen, in denen Merkmalstrukturen notwendigerweise als partiell aufgefaßt werden, zusätzlicher Mechanismen, da prinzipiell nicht zwischen Attributen mit unbekanntem Wert und solchen, die irrelevant für eine Merkmalstruktur sind, unterschieden werden kann. Um zum Beispiel auszudrücken, daß Verben prinzipiell nicht bezüglich Kasus spezifiziert sind, könnte man intuitiv folgenden Constraint über Merkmalstrukturen aufstellen:

$$\neg(\text{CAT} : \mathbf{v} \wedge \text{CASE} : \top) \quad (1)$$

Wobei die Beschreibung \top per Definition von allen Merkmalstrukturen erfüllt wird. Ein Constraint beschreibt die Menge von Merkmalstrukturen, die den Constraint erfüllt. Aus der Perspektive der Informationsverarbeitung wird gefordert, daß Grammatiken deklarativ sind; dazu ist es notwendig, daß das Ergebnis der grammatischen Analyse eines Ausdrucks nicht davon abhängt, in welcher Reihenfolge Constraints wirksam bzw. Unifikationen durchgeführt werden. Um dies zu gewährleisten, muß die Erfüllungsrelation bezüglich der Subsumptionsrelation monoton sein, das heißt, wenn eine Merkmalstruktur \mathcal{A} einen Constraint \mathcal{C} erfüllt und wenn \mathcal{B} eine Merkmalstruktur ist, die von \mathcal{A} subsumiert wird ($\mathcal{A} \sqsubseteq \mathcal{B}$), dann muß auch \mathcal{B} den Constraint \mathcal{C} erfüllen; oder anders ausgedrückt, wenn eine Merkmalstruktur einen Constraint erfüllt, dann muß auch jede spezifischere Merkmalstruktur diesen Constraint erfüllen, da eine Merkmalstruktur die Eigenschaft einen Constraint zu erfüllen monoton erhält und nicht durch die Hinzunahme zusätzlicher Information verlieren kann. Im Bezug auf den Constraint 1 erwarten wir jedoch, daß

$$\left[\begin{array}{l} \text{CAT} : \mathbf{v} \\ \text{PERS} : \mathbf{1} \end{array} \right]$$

den Constraint erfüllt, während

$$\left[\begin{array}{l} \text{CAT} : \mathbf{v} \\ \text{PERS} : \mathbf{1} \\ \text{CASE} : \mathbf{acc} \end{array} \right]$$

den Constraint nicht erfüllen soll. Dies verletzt jedoch die Monotoniebedingung, da

$$\begin{bmatrix} \text{CAT} : & \mathbf{v} \\ \text{PERS} : & 1 \end{bmatrix} \sqsubseteq \begin{bmatrix} \text{CAT} : & \mathbf{v} \\ \text{PERS} : & 1 \\ \text{CASE} : & \mathbf{acc} \end{bmatrix}$$

gilt. Die intuitive Interpretation negierter Constraints ist demnach nicht möglich, wodurch zur Beschränkung der zulässigen Merkmalstrukturen komplexere Mechanismen hinzugezogen werden müssen.¹⁰

Auch unabhängig von dem Problem des Umgangs mit negativen Constraints ist es wünschenswert zwischen irrelevanter und unterspezifizierter Information zu unterscheiden; Pollard and Sag (1987) schreiben hierzu:

“In modelling information it is often important to distinguish between cases where we *lack* information about the value of a certain attribute, and cases where that attribute is *irrelevant* for the kind of object under description.” (siehe Pollard and Sag, 1987, Seite 38)

Eine weitere Schwäche von Merkmalstrukturen ist es, daß die Idee von Kategorien bzw. größeren Klassen linguistischer Objekte nicht inhärent explizit ausgedrückt werden kann. So wird z.B. in dem Lexikon unserer letzten Beispielgrammatik (Seite 3) die kategoriale Information in Form eines zusätzlichen Attributs CAT ausgedrückt. Die zwei lexikalischen Merkmalstrukturen

$$\begin{bmatrix} \text{CAT:} & \mathbf{VP} \\ \text{AGR:} & \begin{bmatrix} \text{NUM} : & \mathbf{sing} \\ \text{PERS:} & 1 \end{bmatrix} \end{bmatrix} \text{ und } \begin{bmatrix} \text{CAT:} & \mathbf{NP} \\ \text{AGR:} & \begin{bmatrix} \text{NUM} : & \mathbf{sing} \\ \text{PERS:} & 2 \end{bmatrix} \end{bmatrix}$$

unterscheiden sich von der Merkmalstruktur

$$\begin{bmatrix} \text{CAT:} & \mathbf{VP} \\ \text{AGR:} & \begin{bmatrix} \text{NUM} : & \mathbf{sing} \\ \text{PERS:} & 2 \end{bmatrix} \end{bmatrix}$$

jeweils nur in einem Attributwert. Dennoch entspricht es unserer Intuition zu sagen, daß die erste Struktur mit der dritten enger verwandt ist, als die zweite mit der dritten, da diese völlig verschieden Arten von Objekten beschreiben. Anders ausgedrückt,

“[...] feature structures come in different *types* depending on the kind of object they describe, and a different set of attributes is appropriate

¹⁰“[...] the presence of negative constraints on features and values cannot be adequately accounted for by the absence of the corresponding positive information. the fact that a given feature structure description does not require that a particular feature label has a particular value cannot be taken to imply that the feature label *does not* have that value.”(Keller, 1993, Seite 20).

for each type of feature structure.” (siehe Pollard and Sag, 1987, Seite 39)

Im weiteren Verlauf dieses Abschnitts werden wir zu getypten Merkmalstrukturen übergehen, also zu Merkmalstrukturen, denen ein expliziter Typ zugewiesen ist. Diese Typen lassen sich ihrerseits wiederum in einer Hierarchie anordnen, so daß es möglich wird auszudrücken, daß eine Merkmalstruktur vom Typ `trans_verb` ist und somit ein transitives Verb beschreibt und da `trans_verb` ein Subtyp des allgemeineren Typs `verb` ist, gilt auch die allgemeinere Aussage, daß die Struktur ein Verb beschreibt.

Ein weiteres Problem bei der Verwendung von Merkmalstrukturen, auf das bereits das kleine Lexikon unserer Beispielgrammatik hinweist, ist, daß Merkmalstrukturen beliebig komplex werden können und der Umfang der lexikalischen Einträge derart zunimmt, daß darunter sowohl die Effizienz der Sprachverarbeitungssysteme leidet, als auch die Verwaltung derartiger Lexika enorme Anstrengungen erfordert:

“Since our theory relies heavily on rich lexical representations, we need to consider what kind of internal organization the lexicon should have. In particular, we do not want to claim that all information contained in lexical entries is simply listed.” (siehe Sag and Wasow, 1999, Seite 173)

Eine Möglichkeit der internen Strukturierung des Lexikons ergibt sich aus der Entwicklung einer Metasprache, in der die Redundanzen im Lexikon erfaßt und in Form von Makros zum Ausdruck gebracht werden können, eine alternative Herangehensweise bietet die Verwendung von getypten Merkmalstrukturen.

1.2.2 Typsignaturen und Wohlgeformtheitskriterien für getypte Merkmalstrukturen

Typsignaturen dienen der Definition einer Menge von wohlgeformten Merkmalstrukturen. Im folgenden soll nun eine informelle Einführung in die Theorie der getypten Merkmalstrukturen geliefert werden.

Der zentrale Gedanke ist, daß allen Merkmalstrukturen ein Typ zugewiesen wird. Diese Typen werden ihrerseits bezüglich ihrer Spezifität in einer Hierarchie unterhalb eines allgemeinsten Typs \top , genannt *Top*, angeordnet. Angereichert wird diese Typhierarchie mit sogenannten *Appropriateness-Bedingungen* (auch *Angemessenheitsbedingungen*), die für alle Merkmalstrukturen eines Typs festlegen, welche Attribute mit Werten von welchem Typ für diese Merkmalstrukturen zulässig sind. Die Entscheidung, die Typen derart anzuordnen, daß spezifischere Typen unterhalb der generelleren zu liegen kommen, ist der Tatsache geschuldet, daß sich diese Anordnung innerhalb der HPSG-Literatur aufgrund der Konvention in den Standardwerken Pollard and Sag (1987, 1994) durchgesetzt hat. Wenn

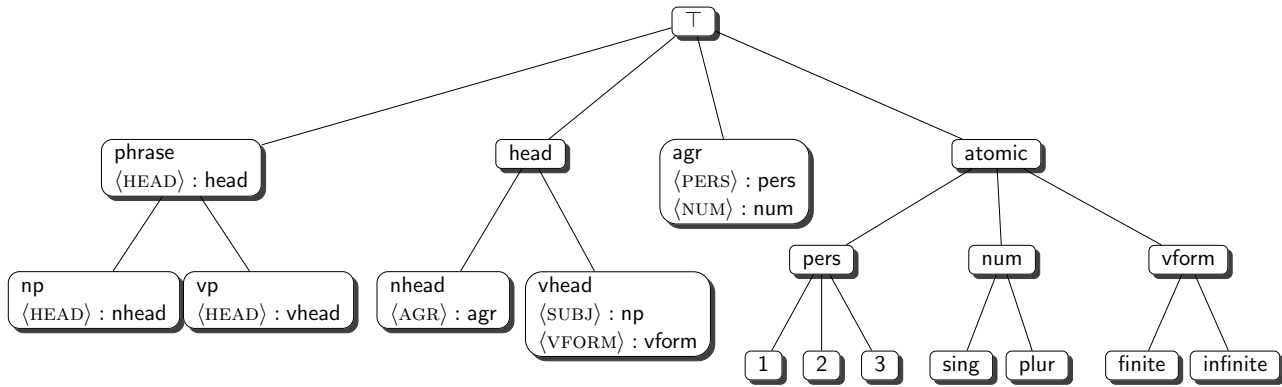


Figure 8: Beispiel einer Typsignatur

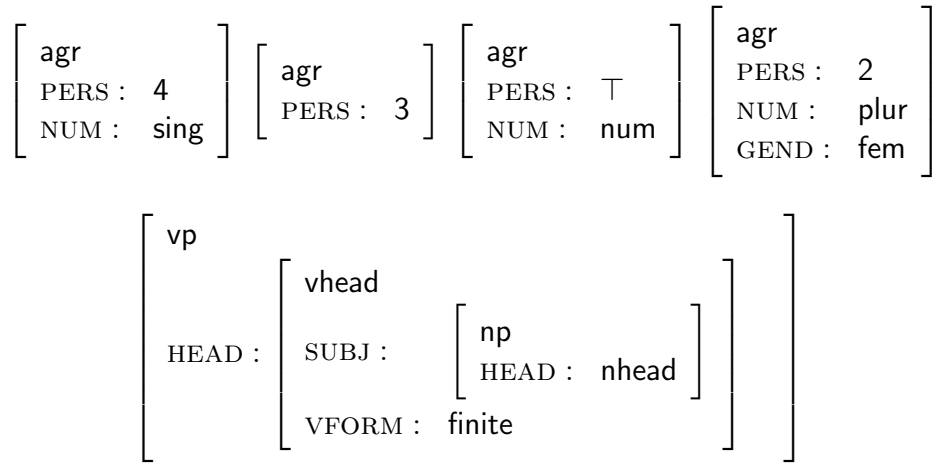


Figure 9: Getypte Merkmalstrukturen, die bezüglich der Typsignatur aus Abbildung 8 nicht vollständig wohlgetypt sind.

ein Typ s jedoch spezifischer ist, als ein Typ t , dann schreiben wir dennoch $t \sqsubseteq s$ und sagen, daß s von t subsumiert wird. Ist \mathcal{T} die Menge aller Typen, dann ist die Typhierarchie also $(\mathcal{T}, \sqsubseteq)$.

Wohlgeformte Merkmalstrukturen, die im Rahmen dieses Ansatzes *vollständig wohlgetypt* genannt werden, sind gerade die Strukturen, die alle Attribute mit den entsprechenden Werten aufweisen, die von den Appropriateness-Bedingungen vorgeschrieben werden und darüberhinaus keine weiteren Attribute besitzen. Ein Beispiel soll dieses Vorgehen verdeutlichen:

Abbildung 8 zeigt eine einfache Typsignatur, bestehend aus 19 Typen, die für die erste Beispielsgrammatik aus Shieber (1986) entwickelt wurde. Auf den Typen ist eine partielle Ordnung definiert, die sogenannte *Subtyprelation* oder *Subsumptionsrelation*. Die Typen bilden zusammen mit dieser Ordnung eine Typhierarchie. In unserem Beispiel hat der Typ *atomic* insgesamt 10 echte *Subtypen*

oder anders ausgedrückt **atomic** ist ein echter *Supertyp* von 10 Typen.¹¹ Wir notieren die Subtyprelation aus Gründen, die im weiteren Verlauf dieses Kapitels deutlich werden, mit dem selben Symbol wie die Subsumptionsrelation ungetypter Merkmalstrukturen; wir schreiben also

atomic \sqsubseteq **plur**.

Für sechs der Typen, nämlich **phrase**, **vp**, **np**, **vhead**, **nhead** und **agr**, sind zusätzlich Appropriateness-Bedingungen definiert, die die Menge der wohlgeformten Merkmalstrukturen beschränken. Die Appropriateness-Bedingung für den Typ **agr** legt zum Beispiel fest, daß Merkmalstrukturen vom Typ **agr** genau zwei Attribute haben, nämlich **PERS** und **NUM**. Bezüglich der Werte der Attribute wird gefordert, daß der Wert von **PERS** eine Merkmalstruktur vom Typ **pers** oder eines Subtyps von **pers** ist; in gleicher Weise sind als Werte für **NUM** nur Merkmalstrukturen vom Typ **num**, **sing** oder **plur** zugelassen. Abbildung 9 zeigt getypte Merkmalstrukturen, die aus unterschiedlichen Gründen bezüglich der Typsignatur aus Abbildung 8 nicht vollständig wohlgetypt sind.¹² Die erste Struktur in der ersten Zeile ist nicht vollständig wohlgetypt, da **4** kein Typ der Typsignatur ist; in der zweiten Struktur fehlt das für Strukturen vom Typ **agr** zwingend vorgeschriebene Attribut **NUM**; die dritte Struktur ist nicht zulässig, da der Typ \top kein Subtyp des Typen **pers** ist; und die vierte Struktur weist ein Attribut **GEND** auf, das nicht durch die Appropriateness-Bedingung für den Typen **agr** legalisiert ist. Die größere Struktur in der zweiten Zeile erscheint auf den ersten Blick vollständig wohlgetypt, jedoch enthält sie die eingebettete Struktur $\left[\text{nhead} \right]$, für die die Typsignatur das Attribut **AGR** vorschreibt. Eine vollständig wohlgetypte Merkmalstruktur vom Typen **vp** zeigt zum Beispiel Abbildung 10.

Im Rahmen der Arbeiten zu getypten Merkmalstrukturen findet sich neben der vollständigen Wohlgetyptheit noch andere Wohlgeformtheitskriterien, die sowohl schwächer als auch stärker als das soeben diskutierte sein können; am häufigsten finden sich Forderungen nach wohlgetypten Merkmalstrukturen, bzw. nach Merkmalstrukturen, die *sort-resolved* sind, was wir hier sinngemäß mit *vollständig*

¹¹In der Regel wird die Subtyprelation als reflexive Relation definiert, dadurch ist **atomic** sowohl ein Subtyp, als auch ein Supertyp von sich selbst. Wir sprechen von echten Subtypen bzw. Supertypen, wenn wir die Verschiedenheit der beteiligten Typen voraussetzen.

¹²Die Notation der getypten Merkmalstrukturen lehnt sich an Carpenter (1992) an: Der Typ einer Merkmalstruktur wird innerhalb der Strukturklammern in der ersten Zeile angegeben. Für die Typenbezeichner werden Kleinbuchstaben und für die Attribute Kapitälchen verwendet. Tritt als Wert eines Attributs eine *atomare Merkmalstruktur* auf, das heißt eine Struktur ohne Attribute, dann werden die Strukturklammern weggelassen. Wir schreiben also

$\left[\begin{array}{l} \text{agr} \\ \text{NUM} : \text{sing} \end{array} \right]$ statt $\left[\begin{array}{l} \text{agr} \\ \text{NUM} : \left[\text{sing} \right] \end{array} \right]$. Wenn wir uns jedoch ausschließlich auf die ato-

mare Merkmalstruktur $\left[\text{sing} \right]$ beziehen wollen, dann verwenden wir die Klammern, um eine Verwechslung mit dem Typen **sing** zu vermeiden.

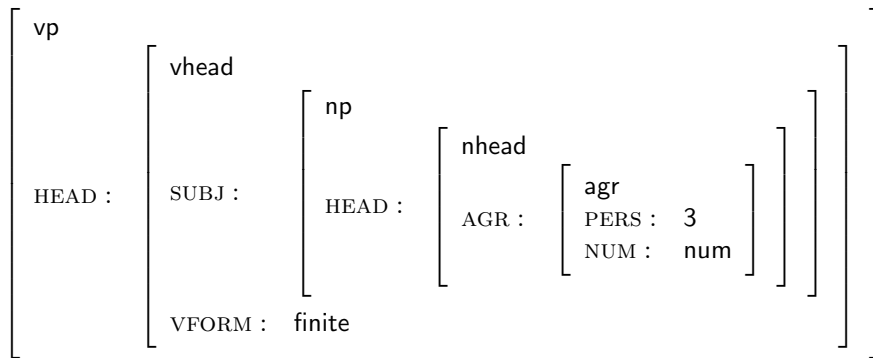


Figure 10: Vollständig wohlgetypte Merkmalstruktur zur Typsignatur aus Abbildung 8

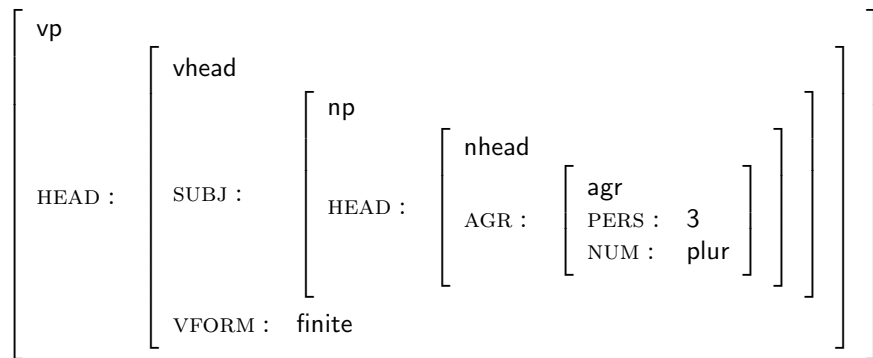


Figure 11: Vollständig ausspezifiziert und vollständig wohlgetypte Merkmalstruktur zur Typsignatur aus Abbildung 8

ausspezifiziert übersetzen werden. Hängt die Wohlgeformtheit einer Merkmalstruktur ausschließlich davon ab, daß sie wohlgetypt ist, dann sind in Abbildung 9 sowohl die zweite Struktur der ersten Zeile als auch die Struktur der zweiten Zeile wohlgeformt. Mit Wohlgetyptheit wird nämlich das Kriterium bezeichnet, das verlangt, daß alle in einer Merkmalstruktur auftretenden Attribute und deren Werte von den Appropriateness-Bedingungen der Typsignatur lizenziert werden. Eine Merkmalstruktur ist vollständig ausspezifiziert, wenn in ihr kein Typ vorkommt, zu dem es in der Typhierarchie einen spezifischeren gibt; in der Regel wird gefordert, daß vollständig ausspezifizierte Merkmalstrukturen auch vollständig wohlgetypt sind. So ist zum Beispiel die Merkmalstruktur 10 nicht vollständig ausspezifiziert, da in ihr die Struktur $\left[\text{num} \right]$ eingebettet ist, wobei *num* kein spezifischster Typ ist; die Typen *sing* und *plur* sind nämlich Subtypen von *num*. Eine vollständig wohlgetypte und vollständig ausspezifizierte Merkmalstruktur zur Typsignatur 8 zeigt Abbildung 11.

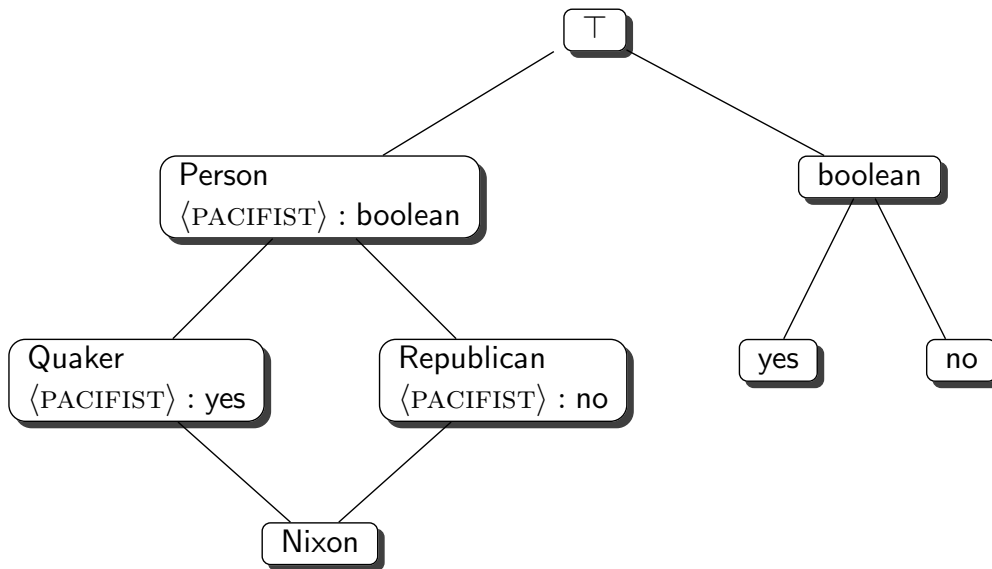


Figure 12: Nixon Diamond: keine wohlgeformte Typsignatur

1.2.3 Typsignaturen sind Vererbungshierarchien

Eine Typsignatur kann als eine Vererbungshierarchie aufgefaßt werden, in der die Appropriateness-Bedingungen eines Typen an seine Subtypen vererbt werden. Diese Sichtweise einer Typsignatur als Vererbungshierarchie begründet auch die zur Subsumptionsrelation duale Darstellung der Hierarchie: In der Literatur zu Vererbungshierarchien (Touretzky, 1986)(???) ist es üblich, daß Information in Hierarchien von oben nach unten vererbt wird.

Da Subtypen die Information ihrer Supertypen erben, muß jede Appropriateness-Bedingung, die an einem Subtypen eingeführt wird, kompatibel mit den Appropriateness-Bedingungen der Supertypen sein. Restringieren zwei Appropriateness-Bedingungen die Werte desselben Attributs, so kann der Attributwert am Subtypen nur verschärft werden. In der Typsignatur aus Abbildung 8 sehen wir ein Beispiel einer solchen Verschärfung: Am Typen `phrase` finden wir die Appropriateness-Bedingung `<HEAD>:head`, diese wird am Typen `vp` verschärft zu `<HEAD>:vhead`. In der Typhierarchie ist `vhead` ein Subtyp von `head`, wäre dies nicht der Fall, dann würde die Vererbungsbedingung der Appropriateness-Bedingungen verletzt und unsere Typsignatur wäre nicht wohlgeformt.

Speziell bei multipler Vererbung, also immer dann, wenn ein Typ Supertypen hat, die untereinander unvergleichbar sind, muß darauf geachtet werden, daß die vererbte Information miteinander kompatibel ist. Abbildung 12 zeigt eine Vererbungshierarchie, die als ‘Nixon Diamond’ bekannt geworden ist und keine wohlgeformte Typsignatur bildet: Merkmalstrukturen vom Typ `Nixon` erben die Information ‘`<PACIFIST>:yes`’ vom Typen `Quaker` und ‘`<PACIFIST>:no`’ vom Typen

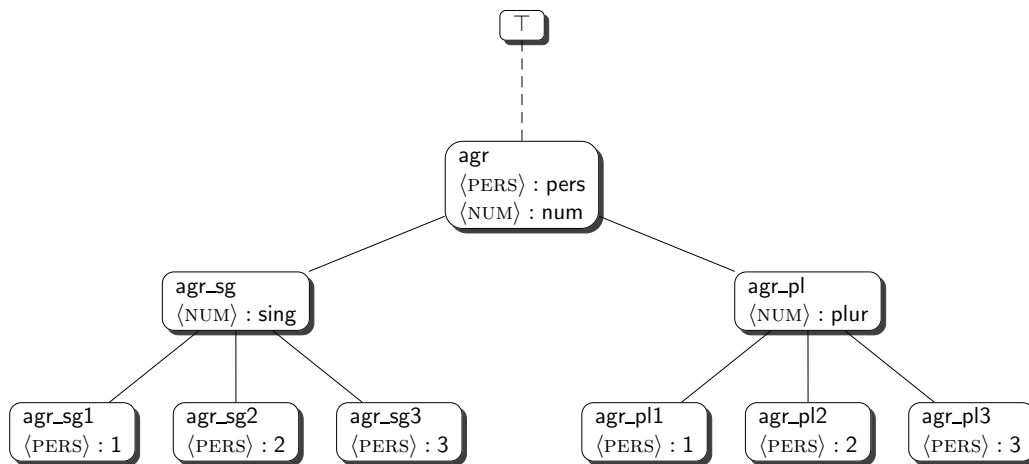


Figure 13: Ausschnitt einer modifizierten Version der Typsignatur aus Abbildung 8.

Republican, die nicht miteinander verträglich ist, da die Typen `yes` und `no` keinen gemeinsamen Subtypen haben.

Ein weiteres Beispiel für die Vererbung der Appropriateness-Bedingungen in einer wohlgeformten Typsignatur zeigt Abbildung 13; es handelt sich hierbei um einen Ausschnitt aus einer Typsignatur, die aus der Abbildung 8 hervorgegangen ist, wobei der Typ `agr` durch die Hinzunahme von Subtypen ausdifferenziert wurde. Abbildung 14 zeigt eine Auswahl von Merkmalstrukturen, an denen noch einmal vorgeführt werden sollen, wie die unterschiedlichen Wohlgeformtheitskriterien im Bezug auf die Typsignatur aus Abbildung 13 greifen. Die Strukturen (a) und (b) sind beide vollständig wohlgetypt. Merkmalstruktur (c) erfüllt zwar die am Typen `agr_sg` eingeführten Appropriateness-Bedingungen, sie verletzt jedoch die Bedingung `<PERS>:pers`, die von dem Supertypen `agr` geerbt wird; somit ist (c) nur wohlgetypt und nicht vollständig wohlgetypt. In der Merkmalstruktur (d) ist die in (c) fehlende Information aufgenommen, so daß (d) vollständig wohlgetypt ist. Merkmalstruktur (e) ist vom Typen `agr_3sg` und erfüllt zwar die Appropriateness-Bedingungen die für `agr_3sg` und den Supertypen `agr` definiert sind, jedoch verletzt es die Bedingung `<NUM>:sing`, die von dem Supertypen `agr_sg` geerbt wird. Somit ist die Merkmalstruktur (e) nicht einmal wohlgetypt. Merkmalstruktur (f) ist schließlich wie Struktur (c) wohlgetypt, aber nicht vollständig wohlgetypt.

1.2.4 Subsumption, Unifikation und Generalisierung

Auch getypte Merkmalstrukturen lassen sich als gerichtete Graphen darstellen, wobei im Gegensatz zu den ungetypten Strukturen jeder Knoten des Graphen beschriftet ist. Abbildung 15 zeigt ein Beispiel einer getypten Merkmalstruktur dargestellt als AVM und als gerichteter (azyklischer) Graph.

In dem Abschnitt über ungetypte Merkmalstrukturen haben wir bereits gese-

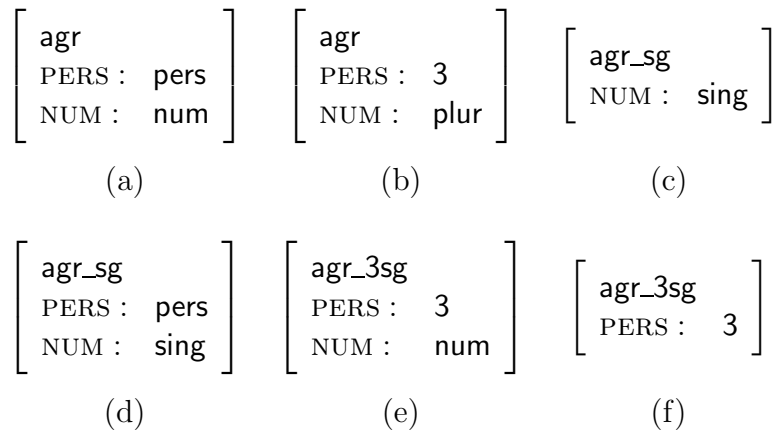


Figure 14: Auswahl von Merkmalstrukturen. Die Strukturen (a), (b) und (d) sind vollständig wohlgetypt bezüglich der Typsignatur aus Abbildung 13.

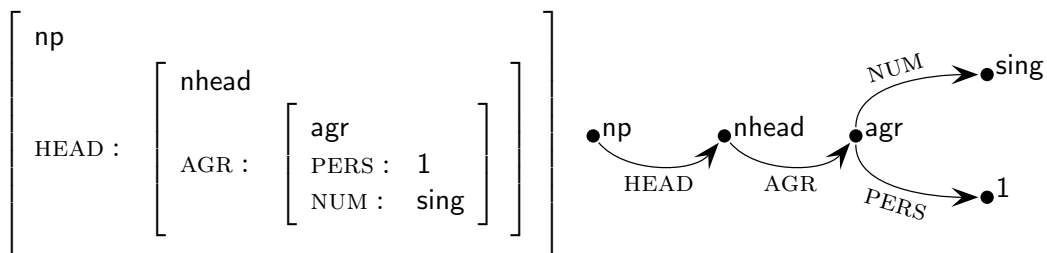


Figure 15: Getypte Merkmalstruktur zu *yo*, dargestellt als AVM (attribute value matrix) und als DAG (directed acyclic graph).

hen, wie diese bezüglich ihrer Spezifität geordnet werden können.¹³ Auch auf den getypten Merkmalstrukturen läßt sich eine solche partielle Ordnung definieren, die ebenfalls Subsumptionsrelation genannt wird. Während es jedoch im ungetypten Fall ausreichend war den Informationsgehalt der Attribut-Wert-Paare zweier Merkmalstrukturen miteinander zu vergleichen, so ist im getypten Fall zu beachten, daß die zugrundeliegende Typsignatur eine zusätzliche Spezifitäts-hierarchie beisteuert. Im Bezug auf die Typsignatur aus Abbildung 13 und die Merkmalstrukturen aus Abbildung 14 ergeben sich unter anderen folgende Subsumptionsbeziehungen: Struktur (a) subsumiert alle anderen Strukturen, bis auf (c) und (f), da diese beiden Merkmalstrukturen zwar von einem spezifischeren Typ als (a) sind, aber ihr Informationsgehalt bezüglich der Attribut-Wert-Paare weniger spezifisch ist, als der von (a). Auf der anderen Seite sind auch

$$\left[\begin{array}{l} \text{agr_3sg} \\ \text{PERS : pers} \\ \text{NUM : num} \end{array} \right] \text{ und } \left[\begin{array}{l} \text{agr_sg} \\ \text{PERS : 3} \\ \text{NUM : sing} \end{array} \right]$$

bezüglich der Subsumptionsrelation unvergleichbar, da die erste Struktur zwar von einem spezifischeren Typ als die zweite ist, diese wiederum hat jedoch bezüglich der Attribut-Wert-Paare den umfassenderen Informationsgehalt.

Die Unifikations- und die Generalisierungsoperation ist analog zum ungetypten Fall definiert. Das Resultat der Unifikation zweier getypter Merkmalstrukturen ist die allgemeinste Struktur, die von beiden Ausgangsstrukturen subsumiert wird. Umgekehrt erhält man durch die Generalisierung die spezifischste Merkmalstruktur, die beide Strukturen subsumiert. Um zu erreichen, daß die Unifikation wenn sie gelingt ein eindeutiges Ergebnis liefert, müssen jedoch zusätzliche Bedingungen an die Typhierarchie gestellt werden. Abbildung 16 zeigt eine Typsignatur, bezüglich der die Unifikation zweier Merkmalstrukturen kein deterministisches Ergebnis liefert: Betrachtet man zum Beispiel die beiden Strukturen

$$\left[\begin{array}{l} \text{agr_sg} \\ \text{NUM : sg} \end{array} \right] \text{ und } \left[\begin{array}{l} \text{agr_fem} \\ \text{GEN : fem} \end{array} \right], \quad (2)$$

so ist die in ihnen enthaltene Information zwar miteinander kompatibel, aber es gibt drei verschiedene Merkmalstrukturen, nämlich

$$\left[\begin{array}{l} \text{agr_sg1fem} \\ \text{NUM : sg} \\ \text{PERS : 1} \\ \text{GEN : fem} \end{array} \right] \text{ und } \left[\begin{array}{l} \text{agr_sg1fem} \\ \text{NUM : sg} \\ \text{PERS : 2} \\ \text{GEN : fem} \end{array} \right] \text{ und } \left[\begin{array}{l} \text{agr_sg1fem} \\ \text{NUM : sg} \\ \text{PERS : 3} \\ \text{GEN : fem} \end{array} \right], \quad (3)$$

¹³Beschränkt man die Menge der Merkmalstrukturen auf die vollständig ausspezifizierten und vollständig wohlgetypten Strukturen, dann lassen sich diese natürlich nicht bezüglich ihrer Spezifität ordnen, da sie ja alle per Definition maximal spezifisch sind. Innerhalb der Menge der (vollständig) wohlgetypten Merkmalstrukturen kommt es jedoch zu einem Informationsgefälle.

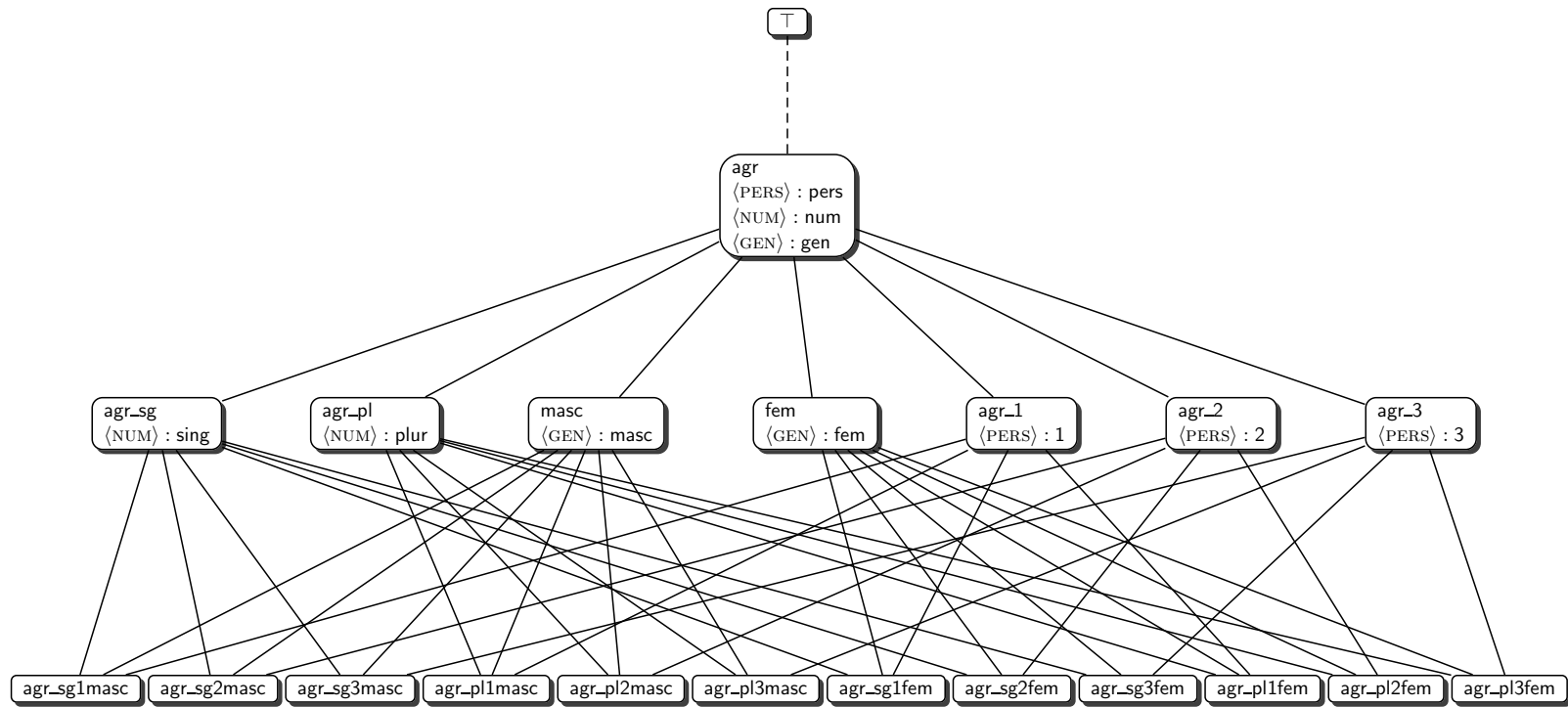


Figure 16: ???

die die Eigenschaft haben, daß sie spezifischer sind als die Strukturen aus 2, wobei es keine Merkmalstruktur mit dieser Eigenschaft gibt, die allgemeiner ist. Alle drei Merkmalstrukturen aus 3 erfüllen somit die Bedingungen des Resultats der Unifikation, was dazu führt, daß die Unifikation zu keinem eindeutigen Ergebnis führt; *die* allgemeinste Merkmalstruktur, die von den Strukturen aus 2 subsumiert wird, existiert nicht.

Diese problematischen Typsignaturen werden ausgeschlossen, indem man per Definition verlangt, daß die Typhierarchie eine sogenannte *BCPO* (bounded complete partial order) bildet, was im endlichen Fall äquivalent ist zu der Forderung, daß zu je zwei Typen genau ein spezifischer Typ existiert, der beide subsumiert. Es ist klar, daß die Hierarchie aus Abbildung 16 diese Bedingung nicht erfüllt, da zum Beispiel die Typen `agr_sg1fem` und `agr_sg2fem` von den Typen `agr_sg` und `agr_fem` subsumiert werden, die in der Hierarchie direkte obere Nachbarn von `agr_sg1fem` und `agr_sg2fem` sind. Abbildung 17 zeigt eine Möglichkeit, die Typsignatur aus Abbildung 16 so zu verändern, daß alle Unifikationen, wenn sie gelingen, zu einem deterministischen Ergebnis führen.

ToDo: unbedingt Beispiel einer Unifikation einfügen, bei der im Resultat Attribute auftreten, die in keiner der Ausgangsstrukturen enthalten sind.

ToDo: etwas über completion-Algorithmus
--

1.2.5 Vorteile getypter Merkmalstrukturen

Im Abschnitt 1.1 über ungetypte Merkmalstrukturen, haben wir verschiedene Probleme aufgezeigt, die sich durch die Verwendung von Typen lösen lassen.

Eine der Schwächen von Formalismen, die ungetypte Merkmalstrukturen verwenden, basiert auf der eingeschränkten Fähigkeit natürliche Klassen linguistischer Objekte darzustellen. Auch wenn es durch die Generalisierungsoperation prinzipiell möglich ist, die Gemeinsamkeiten einer Menge von Merkmalstrukturen auszudrücken, reicht diese Methode nicht aus, um wesentliche Klassenmerkmale herauszuarbeiten. Abbildung 18 zeigt die lexikalische Merkmalstruktur eines intransitiven und eines transitiven Verbs, wie sie Shieber in Shieber (1986) vorschlägt. Die Information über die Subkategorisierung bzw. die Valenz der Verben wird mit Hilfe des Attributs `SUBCAT` zum Ausdruck gebracht, dessen Wert eine Merkmalstruktur ist, die die Liste der vom Verb geforderten syntaktischen Konstituenten repräsentiert. Generalisiert man nun über alle möglichen Merkmalstrukturen transitiver und intransitiver Verben, so erhält man die in Abbildung 19 gezeigten Strukturen, die demnach die Klasse der Verben und die Teilklassen der intransitiven und der transitiven Verben repräsentieren müßten. Natürlich tun dies diese Strukturen in gewisser Weise auch, sie sind jedoch nicht

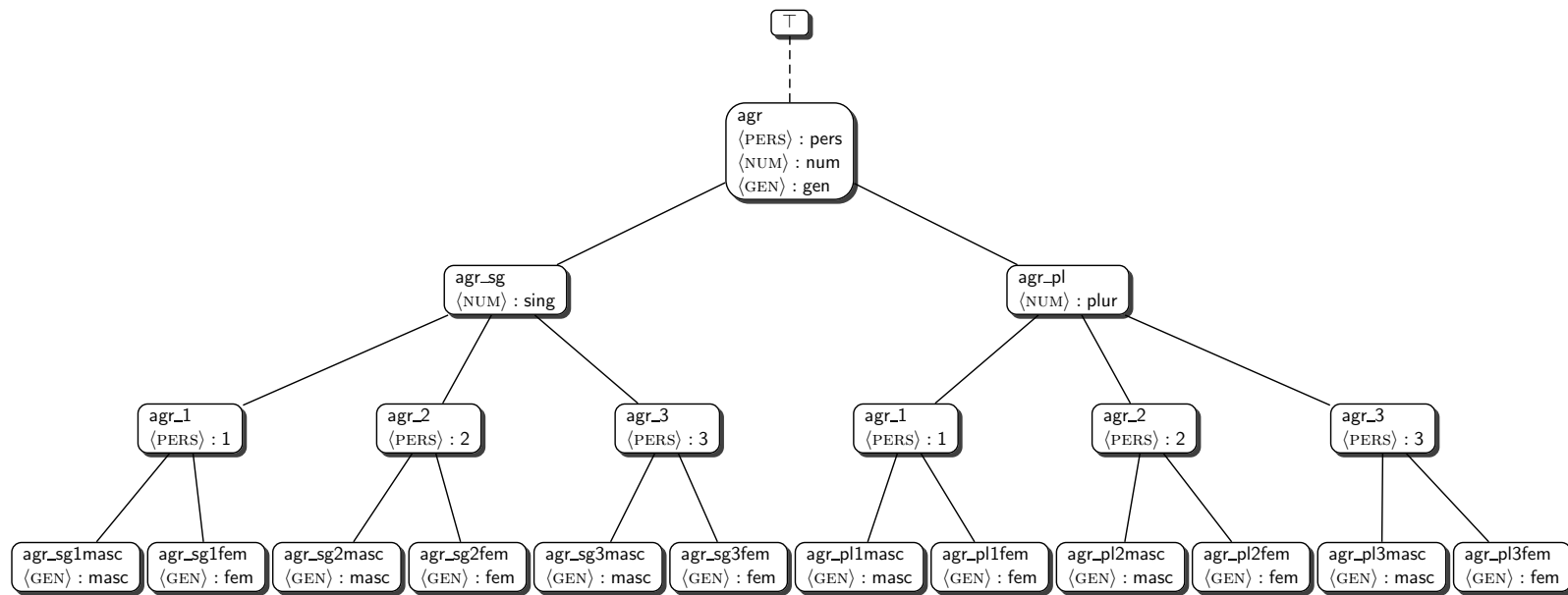


Figure 17: ???

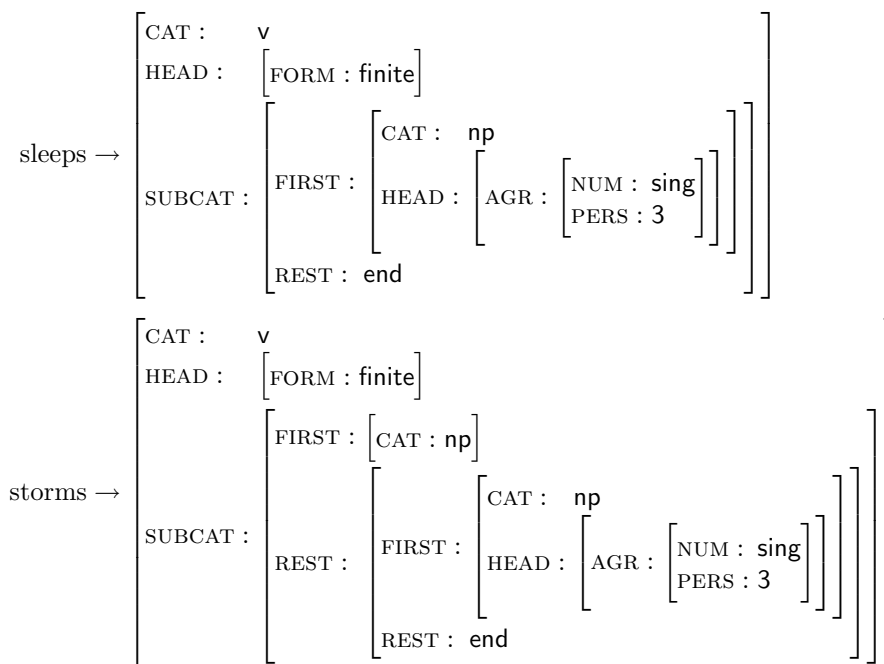


Figure 18: Ungetypte Merkmalstruktur eines intransitiven und eines transitiven Verbs mit Subkategorisierungsrahmen (nach Shieber, 1986).

geeignet in einem Grammatikformalismus als ‘allgemeines Verb’, ‘allgemeines intransitives Verb’ oder ‘allgemeines transitives Verb’ zu agieren: Die Struktur ‘allgemeines Verb’ ist zu unspezifisch, ihr fehlt jegliche Information darüber, daß im Englischen das Verb eines Satzes mit dem Subjekt kongruiert. In dem Wert des Attributs SUBCAT kommt das Subjekt nicht vor, es wird lediglich spezifiziert, daß die dem Verb am engsten verbundene subkategorisierten Konstituente eine Nominalphrase sein muß und somit das Attribut-Wert-Paar CAT:np umfaßt.¹⁴ Betrachtet man nur den Wert des Attributs SUBCAT, so ist ihm nicht zu entnehmen, daß SUBCAT eine *Liste* der subkategorisierten Konstituenten spezifiziert. Ein anderes Phänomen tritt bei den Merkmalstrukturen ‘intransitive Verb’ und ‘transitives Verb’ auf, da diese sehr komplex sind. Diese Strukturen sind zwar spezifisch genug, jedoch sind die wesentlichen Kennzeichen zur Identifizierung der Verbklassen, nämlich SUBCAT : REST : end für intransitive Verben und SUBCAT : REST : REST : end für transitive Verben, sehr tief eingebettet, was ihre Verwendung in einer Grammatik umständlich macht. Die Einführung zweier neuer Werte *intrans_verb* und *trans_verb* stellt auch keinen Ausweg aus der Situation dar, da dadurch die Struktur ‘allgemeines Verb’ nicht mehr allgemeiner als die Strukturen der Verbteilklassen wäre.

¹⁴Hier wurde außen vor gelassen, daß Verben auch andere Phrasen, z.B. Präpositionalphrasen, subkategorisieren können.

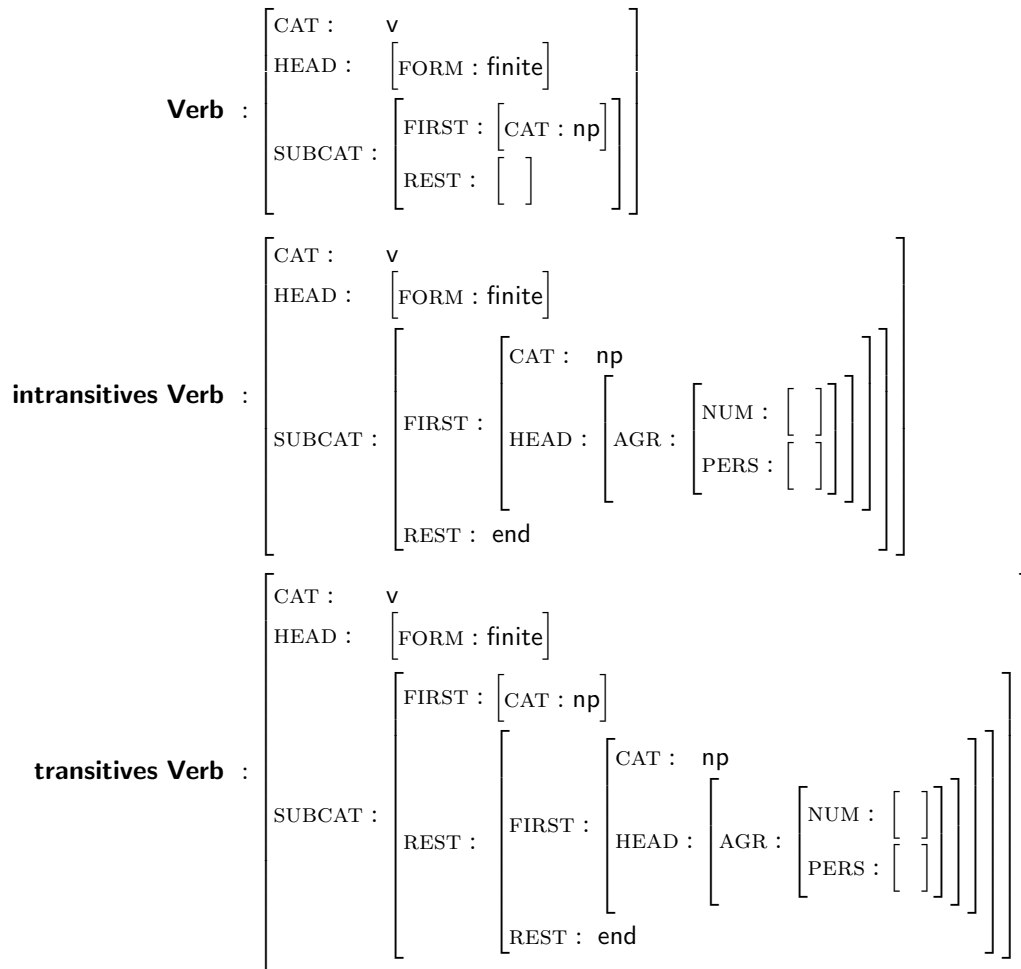


Figure 19: ??? 18.

Die Verwendung von hierarchisch geordneten Typen löst dieses Problem, da Typen eingesetzt werden können, um Merkmalstrukturen transparent in natürliche Klassen zu klassifizieren. Teilklassenbeziehungen entsprechen den Subtypbeziehungen in der Typhierarchie. Die Anforderungen, die an die Appropriateness-Bedingungen gestellt werden, garantieren, daß die gesamte Information, die eine Merkmalstruktur eines Typs ausdrückt, immer auch von allen Merkmalstrukturen von Subtypen des Typen umfaßt wird. Um Listen, wie die Subkategorisierungslisten auszudrücken, bietet sich eine Typsignatur wie in Abbildung 20 an, die darüberhinaus auch die allgemeinste vollständig wohlgetypte Merkmalstruktur einer nicht leeren Liste vom Typen `ne_list` zeigt. Um zum Beispiel die Valenzen der verschiedenen Verbklassen angemessen darzustellen, ist eine Typsignatur wie in Abbildung 22 denkbar, in der einzelne Typen für 1-, 2- bzw. 3-stellige Subkategorisierungslisten existieren.

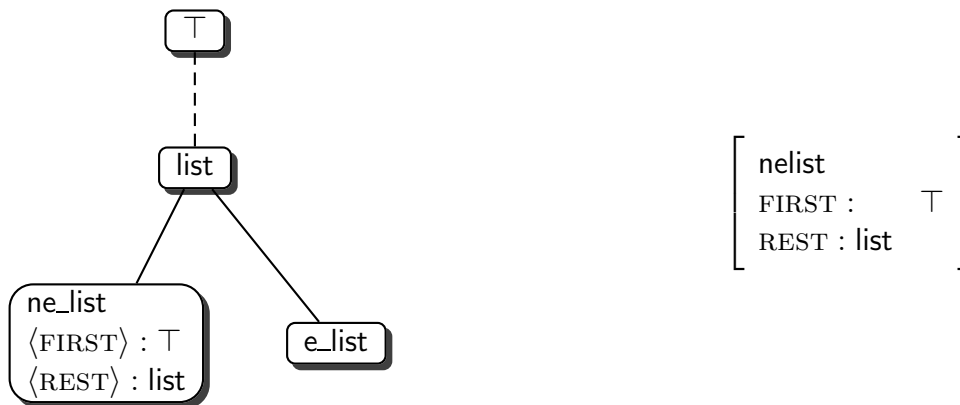


Figure 20: Typsignatur zur Kodierung von Listen. Der Typ `ne_list` bezeichnet die Klasse der nicht-leeren Listen (non empty list) und `e_list` die Klasse der leeren Listen (empty list).

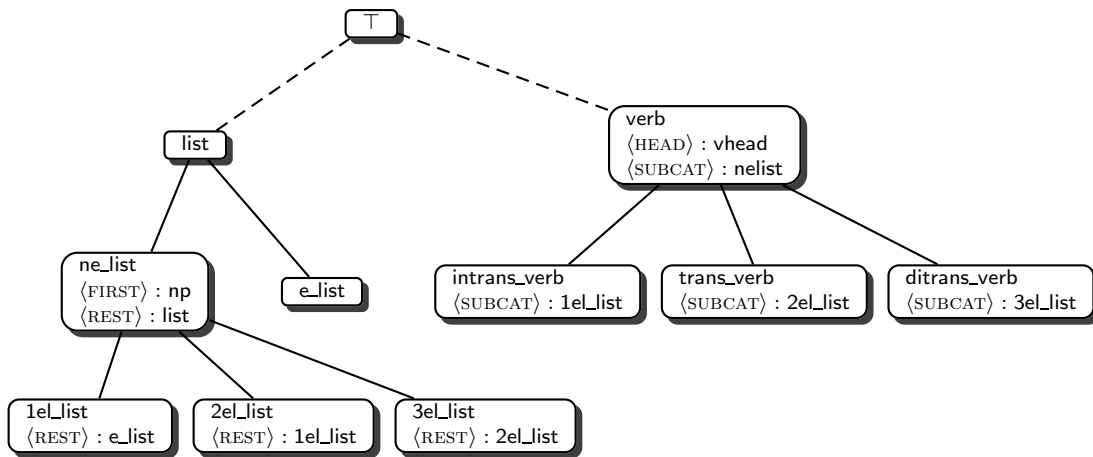


Figure 21: Ausschnitt einer Typsignatur, die die Subkategorisierungen verschiedener Verbklassen erfaßt.

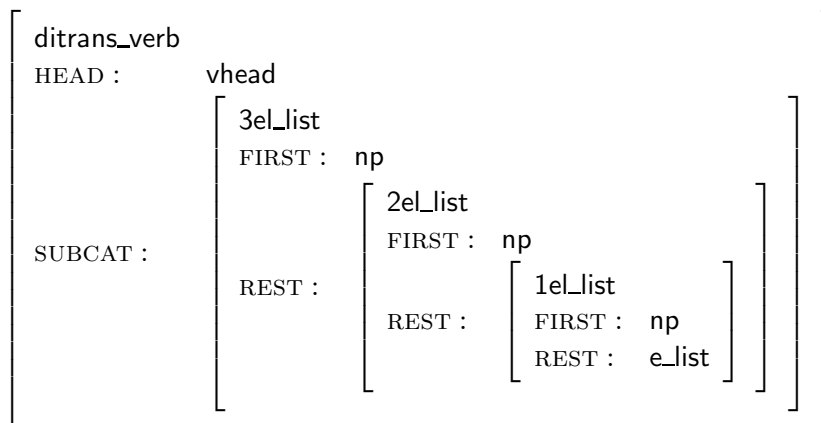


Figure 22: Vollständig wohlgetypte Merkmalstruktur vom Typ `ditrans_verb` zur Typsignatur aus Abbildung 22.

Das zweite Problem, mit dem Formalismen, die auf ungetypten Merkmalstrukturen basieren, nicht umgehen können, ist die Unterscheidung zwischen Attributen, die für eine Merkmalstruktur nicht spezifiziert sind und solchen die irrelevant bzw. nicht angemessen sind. Auch hier bieten Typsignaturen einen direkten Ausweg, da in ihnen festgelegt wird, welches die für Merkmalstrukturen eines bestimmten Typs angemessenen Attribute sind und von welcher Art die Attributwerte sind. Alle Attribute, die von der Typsignatur für einen bestimmten Typen nicht lizenziert sind, können in wohlgetypten Merkmalstrukturen dieses Typs nicht auftreten; sie sind per Definition für diese Strukturen irrelevant. Der Einsatz von Supertypen in der Typhierarchie ermöglicht es, die Werte eines Attributs in einer vollständig wohlgetypten Merkmalstruktur unterspezifiziert zu lassen, obwohl diese Strukturen auszeichnet, daß in ihnen alle angemessenen Attribute auftauchen. So ist zum Beispiel in der vollständig wohlgetypte Merkmalstruktur

$$\left[\begin{array}{l} \text{np} \\ \\ \text{HEAD :} \end{array} \left[\begin{array}{l} \text{nhead} \\ \\ \text{AGR :} \end{array} \left[\begin{array}{l} \text{agr} \\ \text{PERS : } 3 \\ \text{NUM : } \text{num} \end{array} \right] \right] \right]$$

zur Typsignatur 9 der Wert des Attributs NUM in der eingebetteten Struktur vom Typen *agr* unterspezifiziert; durch den Verweis auf den Typen *num* wird jedoch bereits ein angemessener Wertebereich festgelegt. Verschärft man den Wert von NUM nun auf *sing*, so erhält man eine vollständig ausspezifizierte Merkmalstruktur.

Keller (1993) faßt die zusätzlichen Möglichkeiten, die der Einsatz von Typen in unifikationsbasierten Formalismen bietet, wie folgt zusammen:

“The types form a multiple inheritance hierarchy allowing information about grammatical objects to be encoded at different levels of abstraction. In addition, Carpenter’s type system employs *appropriateness conditions* for features, which are used to model the distinction between features whose values are unknown, and those which are simply irrelevant for a given type.” (siehe Keller, 1993, Seite 34)

Im Rahmen unserer kleinen Beispielsgrammatik des Spanischen in Abschnitt 1.1 haben wir auf das Problem der enorm anwachsenden lexikalischen Information bei unifikationsbasierten Formalismen hingewiesen. Ein großer Teil der den lexikalischen Objekten zugeordneten Information ist redundant und sollte, auch um die Lexika leichter warten zu können und somit die Fehleranfälligkeit zu reduzieren, als Regularitäten erfaßt werden. Typsignaturen ermöglichen eine solche Redundanzvermeidung:

“Structuring the lexicon in terms of an inheritance hierarchy of types has made it possible to factor out information common to many lexical entries, thereby greatly reducing lexical redundancy.” (siehe Sag and Wasow, 1999, Seite 202)

Durch die Ausdifferenzierung des allgemeinen Typs *agr*, wie in der Typsignatur in Abbildung 13, reduziert sich unser Beispielllexikon von Seite 4 zu:¹⁵

<i>yo</i>	↦	np & HEAD AGR: <i>agr_sg1</i>
<i>tú</i>	↦	np & HEAD AGR: <i>agr_sg2</i>
<i>ella</i>	↦	np & HEAD AGR: <i>agr_sg3</i>
<i>nosotros</i>	↦	np & HEAD AGR: <i>agr_pl1</i>
<i>vosotros</i>	↦	np & HEAD AGR: <i>agr_pl2</i>
<i>ellas</i>	↦	np & HEAD AGR: <i>agr_pl3</i>
<i>hablo</i>	↦	vp & SUBJ HEAD AGR: <i>agr_sg1</i> & VFORM: <i>finite</i>
<i>hablas</i>	↦	vp & SUBJ HEAD AGR: <i>agr_sg2</i> & VFORM: <i>finite</i>
<i>habla</i>	↦	vp & SUBJ HEAD AGR: <i>agr_sg3</i> & VFORM: <i>finite</i>
<i>hablamos</i>	↦	vp & SUBJ HEAD AGR: <i>agr_pl1</i> & VFORM: <i>finite</i>
<i>habláis</i>	↦	vp & SUBJ HEAD AGR: <i>agr_pl2</i> & VFORM: <i>finite</i>
<i>hablan</i>	↦	vp & SUBJ HEAD AGR: <i>agr_pl3</i> & VFORM: <i>finite</i>

Ziel des folgenden Abschnittes ist es zu motivieren, warum die *automatische* Transformation von existierenden Lexika mit ungetypten Merkmalstrukturen in getypte Varianten mit entsprechenden, automatisch erstellten, Typsignaturen interessant ist.

References

Lawrence W. Barsalou. Frames, concepts, and conceptual fields. In Adrienne Lehrer and Eva Feder Kittay, editors, *Frames, Fields, and Contrasts*, pages 21–74. Lawrence Erlbaum Associates Publishers, Hillsdale, NJ, 1992.

Leonard Bloomfield. *Language*. Holt, Rinehart and Winston, New York, 1933.

Joan Bresnan. *The Mental Representation of Grammatical Relations*. The MIT Press, Cambridge, MA, 1982.

Bob Carpenter. *The Logic of Typed Feature Structures*. Cambridge Tracts in Theoretical Computer Science 32. Cambridge University Press, 1992.

Noam Chomsky. *Aspects of a Theory of Syntax*. MIT Press, Cambridge, Mass., 1965.

¹⁵Es ist klar, daß man in einem realistischen Lexikon kein Wort als Nominal- oder Verbalphrase analysieren würde, sondern zusätzliche Regeln für die Bildung derartiger Phrasen hätte.

- Noam Chomsky and Morris Halle. *The Sound Pattern of English*. Harper and Row, New York, 1968.
- Gerald Gazdar, E. Klein, Geoff Pullum, and Ivan Sag. *Generalized Phrase Structure Grammar*. Blackwell, Oxford, 1985.
- Roman Jakobson and Morris Halle. *Fundamentals of Language*. 2. überarb. Aufl. 1975, The Hague, 1956.
- Ronald M. Kaplan and Joan Bresnan. Introduction: grammars as mental representations of language. In Bresnan (1982), pages xvii–lii.
- R. T. Kasper and W. C. Rounds. The logic of unification in grammar. *Linguistics and Philosophy*, 13(1):35–58, 1990.
- Bill Keller. *Feature Logics, Infinitary Descriptions and Grammar*. CSLI Lecture Notes 44, Stanford, 1993.
- Michael Kenstowicz and Charles Kisseberth. *Generative Phonology. Description and Theory*. Academic Press Inc., London, 1979.
- Gerald Penn. *The Algebraic Structure of Attributed Type Signatures*. PhD thesis, School of Computer Science, Carnegie Mellon University, 2000.
- C. Pollard and I. Sag. Information-based syntax and semantics. Technical Report 13, Center for the Study of Language and Information, Stanford University, 1987.
- Carl Pollard and Ivan A. Sag. *Head-Driven Phrase Structure Grammar*. The University of Chicago Press, Chicago, 1994.
- Ivan Sag and Thomas A. Wasow. *Syntactic Theory: A Formal Introduction*. Center for the Study of Language and Information, Stanford, California, 1999.
- Stuart M. Shieber. The formalism and implementation of PATR-II. *Research of interactive acquisition and use of knowledge. SRI Menlo Park*, pages 39–79, 1983.
- Stuart M. Shieber. *An Introduction to Unification-Based Approaches to Grammar*. CSLI Lecture Notes 4, Stanford, 1986.
- D. S. Touretzky. *The Mathematics of Inheritance Systems*. Research Notes in Artificial Intelligence. Morgan Kaufmann Publishers, Los Altos, 1986.
- N. Trubetzkoy. *Grundzüge der Phonologie*. 4. Aufl., Göttingen, 1939.