Eine umfassende Constraint-Dependenz-Grammatik des Deutschen

Kilian A. Foth

16. Januar 2006

Inhaltsverzeichnis

Kapitel 1

Das Dependenzmodell

Dies ist eine umfassende Constraint-Dependenz-Grammatik des Deutschen. Sie soll im Prinzip beliebigen Text abdecken können, setzt allerdings aus mehreren Gründen geschriebenen Input voraus:

1. Einige Regeln verwenden die An- und Abwesenheit von Satzzeichen, um bestimmte Konstruktionen zu erlauben. Ohne Satzzeichen sind diese Regeln wirkungslos oder gar hinderlich.

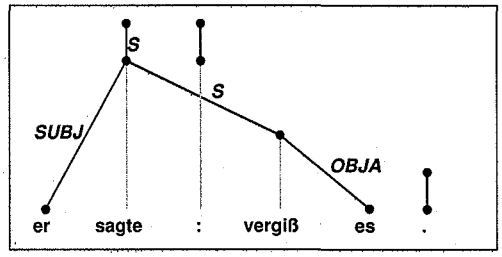
2. Eindeutige sprachliche Fehler werden insofern robust verarbeitet, daß es ungeachtet der Eingabe immer mindestens eine mögliche Struktur geben sollte. Jedoch ist dies nicht immer diejenige, die der Sprechintention entsprechen würde. Beispielsweise wird ein Kongruenzfehler gewöhnlich nicht zu einer veränderten syntaktischen Struktur führen, ein kategorieändernder Schreibfehler aber sehr wohl ("Die Firma will mehr Rechner *verkaufe*"). Daher wird Spontansprache gewöhnlich nicht optimal modelliert.

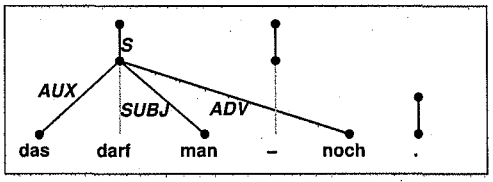
1.1 Annotationsebenen

1.1.1 Die Syntaxebene

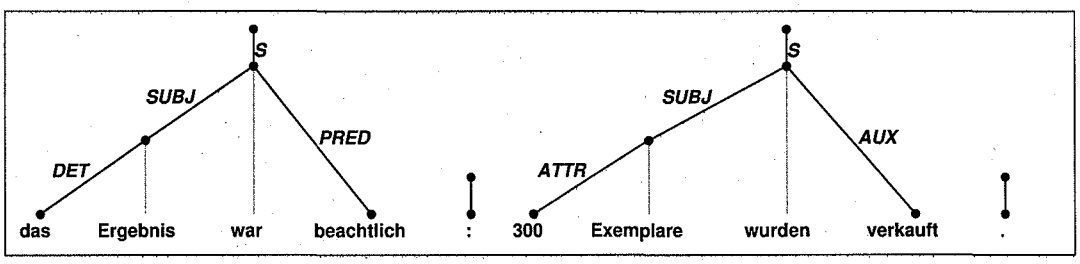
Die Ebene `SYN` stellt die eigentliche Modellierung der Satzstruktur dar. Die anderen Ebenen dienen dazu, Verbindungen zwischen solchen Worten zu markieren, die nicht direkt syntaktisch untergeordnet sind, aber dem Parser dabei helfen, zwischen sonst ununterscheidbaren Strukturen zu wählen. Zum Beispiel darf ein Relativsatz (label `REL`) nur auftreten, wenn er ein Relativpronomen enthält; das Relativpronomen ist aber nicht immer direkt mit dem Nebensatzverb verbunden, daher muß diese Beziehung außerhalb des Syntaxbaumes markiert werden.

Die syntaktische Struktur wird als eine Analyseebene `SYN` aufgebaut. Im allgemeinen sollte eine durch Punkt abgeschlossene Äußerung durch einen zusammenhängenden Baum dargestellt werden. Es gibt jedoch Ausnahmen; die Satzzeichen `:`, `-`, `()` und `;` werden manchmal zur Unterteilung eines vollständigen Satzes verwendet:





Manchmal aber dienen sie zur Abgrenzung zwischen zwei vollständigen Sätzen. In diesem Fall ist es angebracht, zwei Syntaxbäume aufzustellen:



Constraints erlauben daher ausdrücklich mehrere Hauptsätze ohne Bestrafung, wenn sie durch solche Satzzeichen getrennt sind.

Alle Dependenzkanten der Syntaxebene tragen ein erklärendes Label. Es wird grob unterschieden zwischen

- Komplementen: Das sind alle Arten von Wörtern, die mehr oder weniger regelmäßig mit ihrem Regenten auftreten: Subjekte, Artikel, Objekte etc.
- Modifikatoren: Das sind alle anderen Beziehungen.

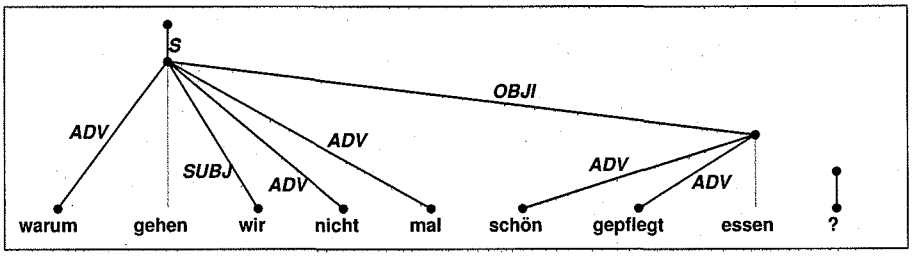
Komplemente werden durch folgende Label ausgedrückt:

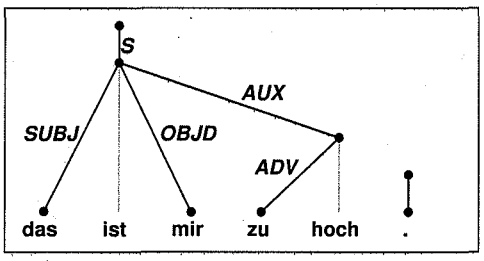
`AUX AVZ CJ DET PN PRED OBJA OBJA2 OBJC OBJD OBJG OBJI OBJP SUBJ`

Alle anderen Label sind Modifikatoren, d.h. niemals obligatorisch.

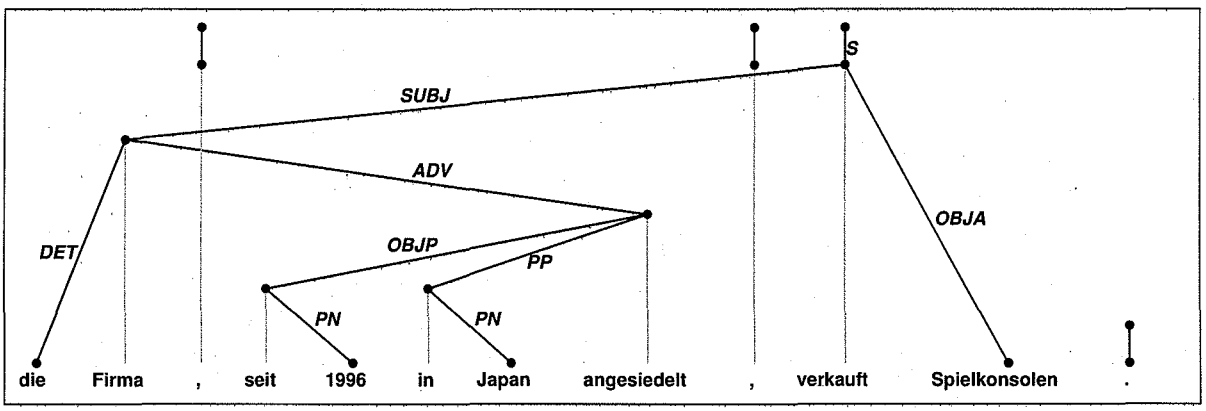
Das Label ADV

`ADV` bezeichnet adverbiale Modifikation, entweder durch wirkliche Adverben (ADV) oder durch verwandte Wortklassen wie ADJD, PTKNEG, VVPP, PWAV oder PTKA.





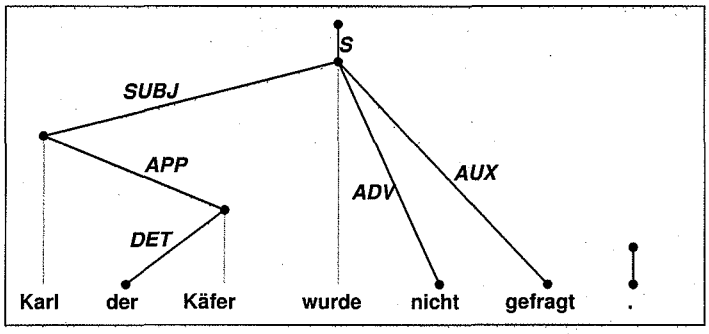
Als ADV wird auch eine nachgestellte Modifikation von Nomen durch prädikatives Adjektiv bezeichnet:

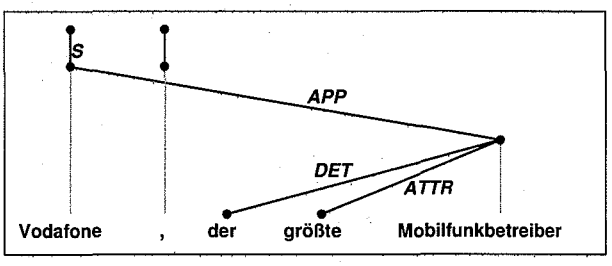


Inhaltlich könnte diese Konstruktion auch als elliptischer Relativsatz angesehen werden; da kein Relativpronomen auftritt, betrachten wir sie aber als Art der adverbialen Bestimmung.

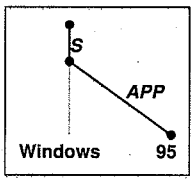
Das Label APP

Das Label APP verbindet aufeinanderfolgende Worte derselben NP, falls sie nicht Determiner oder Attribute sind (Artikel, attributive Pronomen, attributiv gebrauchte Zahlen etc.). Der Normalfall ist die Beziehung zwischen zwei Nomen (NN, NE oder FM), also eine echte Apposition:

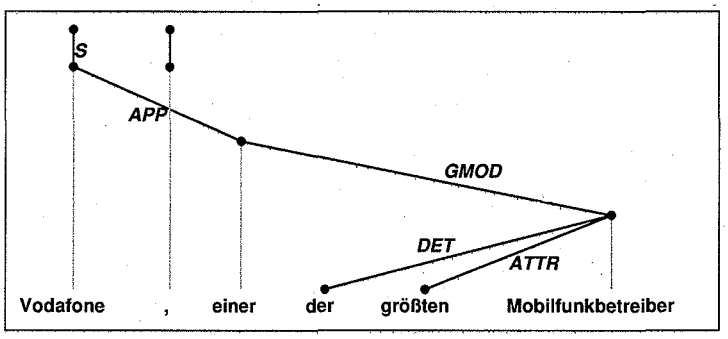




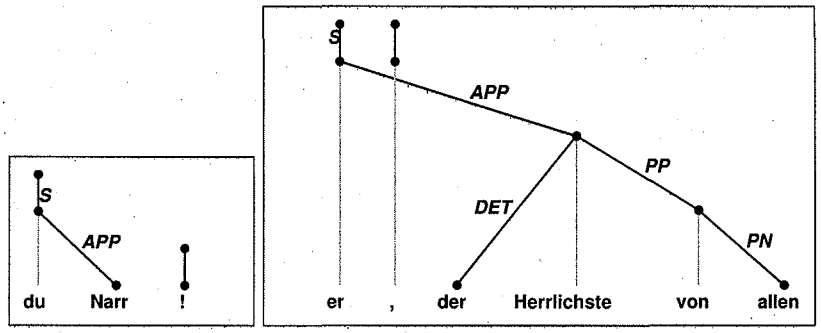
Auch andere aufeinanderfolgende Bestandteile einer Nominalphrase werden mit APP bezeichnet, zum Beispiel nachgestellte Zahlen:



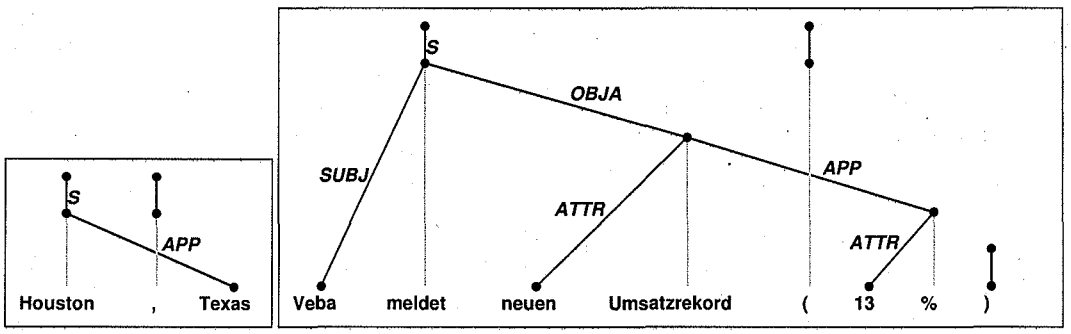
Schließlich können auch substituierende Pronomen als Appositionen auftreten:



Das modifizierte Wort ist normalerweise ein Nomen. Nur ganz selten werden Personalpronomen modifiziert:

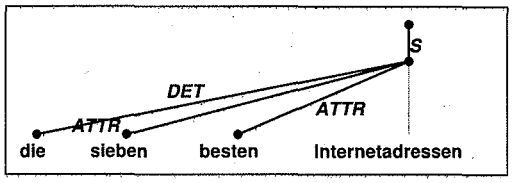


Es gibt eine Reihe von erläuternden Beziehungen zwischen Nominalphrasen, die nicht im strengen Sinne Appositionen (d.h. Gleichsetzungen) sind, sondern nur enge Beziehungen (z.B. 'A hat B', 'A befindet sich am Ort B'). Solche Beziehungen dürfen ebenfalls als APP modelliert werden, wenn sonst keine Unterordnung möglich ist:

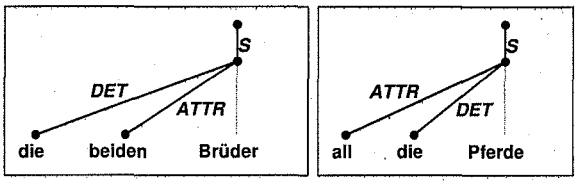


Das Label ATTR

Dies Label bezeichnet Attribute von Nomen. Meistens sind das attributive Adjektive oder Zahlen:



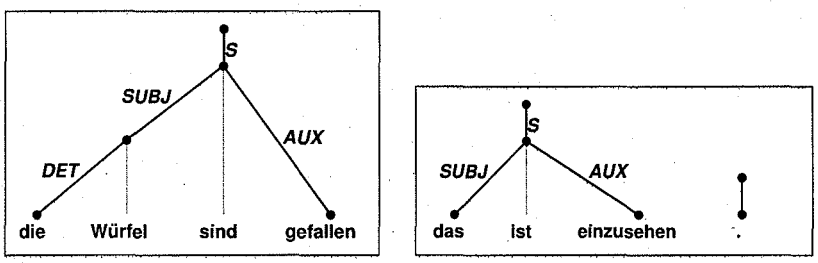
Wenn ein attributives Pronomen (PIDAT) zusammen mit einem Artikel o.ä. steht, dann ist dieser DET und das Pronomen ATTR:



Das Label AUX

Mit AUX werden Verbgruppen zusammengefügt, die aus Hilfsverb und Vollverb bestehen. Grundsätzlich ist immer das finite Verb der Kopf der gesamten Verbgruppe; in nicht-finiten Verbphrasen ist das Vollverb dem Hilfsverb untergeordnet.

Das Verb 'sein' kann viele verschiedene Wortarten als Komplement nehmen: Nomen, Infinitive, Präpositionen, Adverben etc. Wenn es sich um ein Verb handelt, wird es mit AUX bezeichnet:

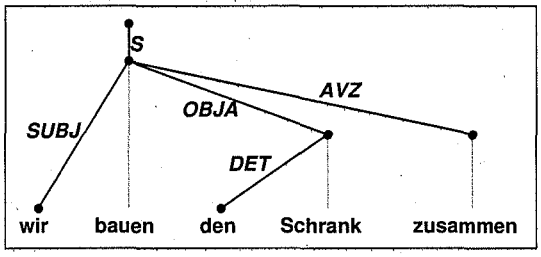


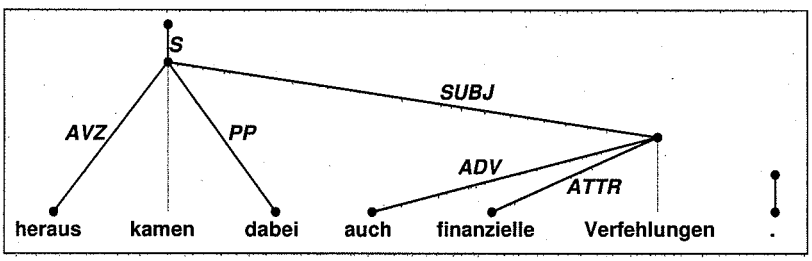
Anderenfalls handelt es sich um ein Prädikat, das mit PRED bezeichnet wird (vgl. unten).

Das Verb 'haben' kann einerseits als Auxiliarverb Verbphrasen bilden, andererseits ein normales Akkusativobjekt (OBJA) tragen.

Das Label AVZ

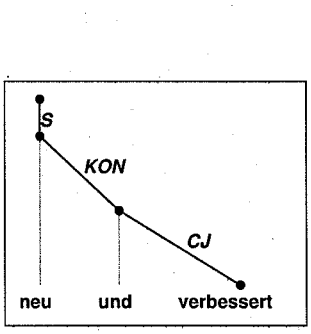
Diese Kante verbindet ein abgetrenntes Verb-Präfix mit seinem finiten Verb. Steht das Präfix alleine, so muß es dem Verb folgen oder unmittelbar vorausgehen.

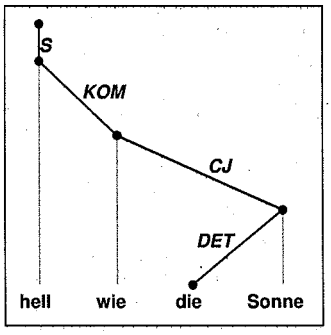


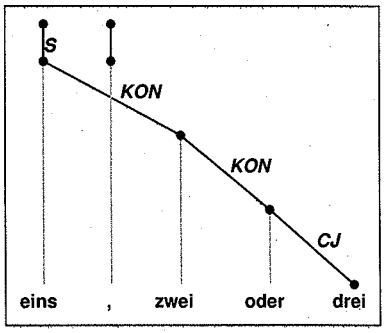


Das Label CJ

Hiermit wird das Komplement einer beiordnenden oder vergleichenden Konjunktion bezeichnet, also z.B. von "und", "oder", "wie" und "als". Es kann vielen verschiedenen Wortklassen angehören.







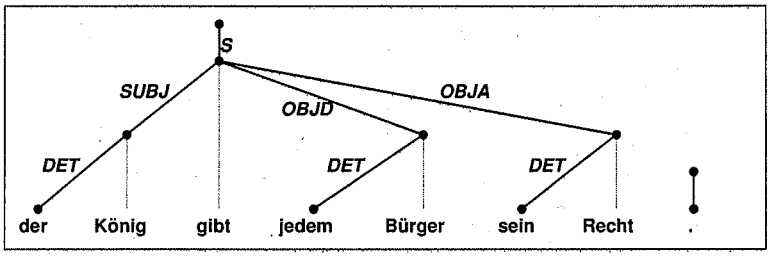
Das Komplement wird immer mit `CJ` bezeichnet, egal welche logische Funktion es erfüllt.

Einige Konjunktionen können mehr als zwei Elemente beiordnen. In diesem Fall wird nur das letzte mit `CJ` bezeichnet, alle anderen mit `KON`.

Zur Struktur von koordinierten Phrasen vgl. das Label `KON`.

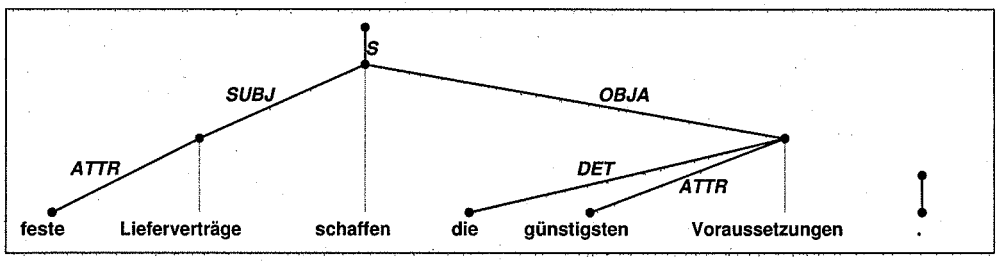
Das Label DET

Dieses Label bezeichnet den Determiner eines Nomen. Das sind gewöhnlich bestimmte oder unbestimmte Artikel, aber auch alle Arten von attributiven Pronomen:

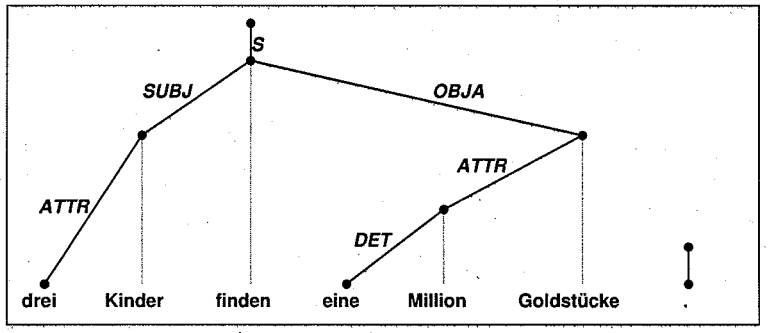


Artikel sind immer mit `DET` zu bezeichnen, außer sie sind beigeordnet (dann `CJ`) oder fragmentarisch (dann `S`).

Adjektive sind niemals `DET`, sondern immer `ATTR`:

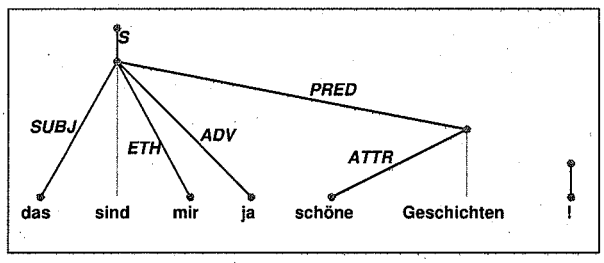


Normale Zahlen (CARD) und Zahlen, die Nomina sind, werden wie Adjektive mit ATTR bezeichnet.



Das Label ETH

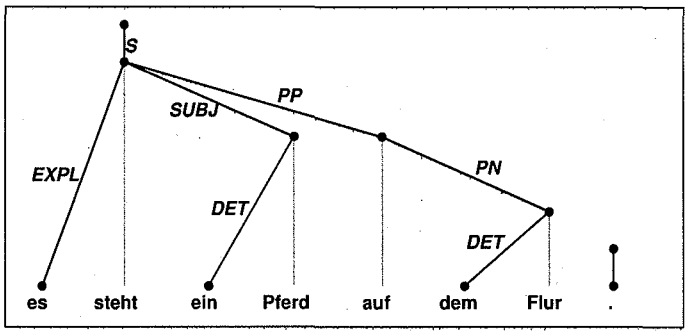
Dies Label bezeichnet Dativunterordnungen, die nicht einem normalen Verbrahmen entsprechen, sondern frei eintreten:

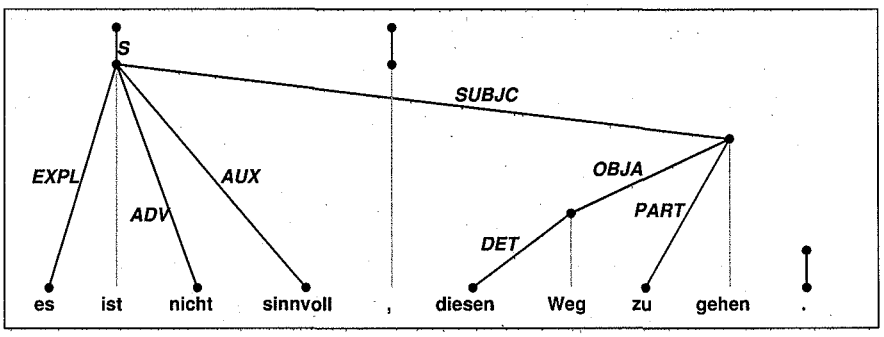


Zur Unterscheidung zwischen ETH und OBJD siehe Abschnitt 1.2.4.

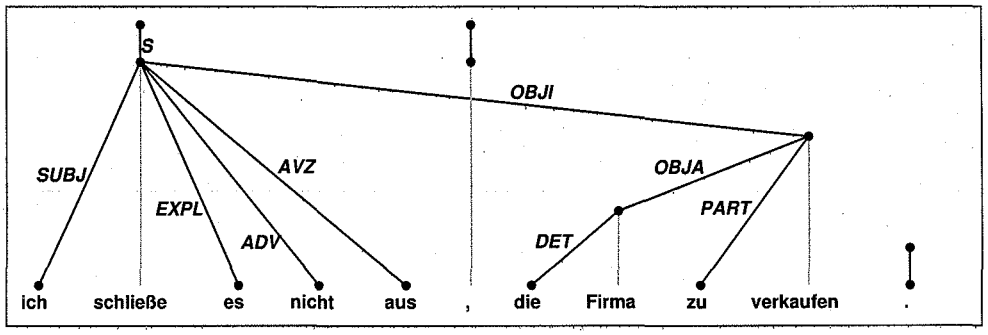
Das Label EXPL

Dieses Label bezeichnet das expletive 'es' an einem finiten Verb. Es tritt in zwei Varianten auf. Im Hauptsatz steht es vor dem Verb, wenn das Subjekt nachgestellt wird und keine andere Konstituente das Vorfeld füllen kann. (Das nachgestellte Subjekt kann auch ein Subjektsatz sein.)

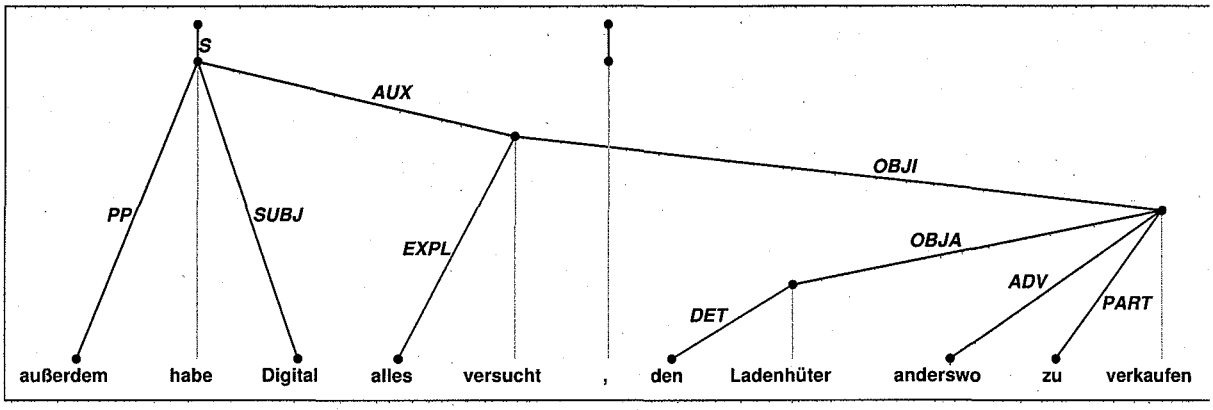




Nach dem Verb kann es als Vertreter eines Objektsatzes stehen:



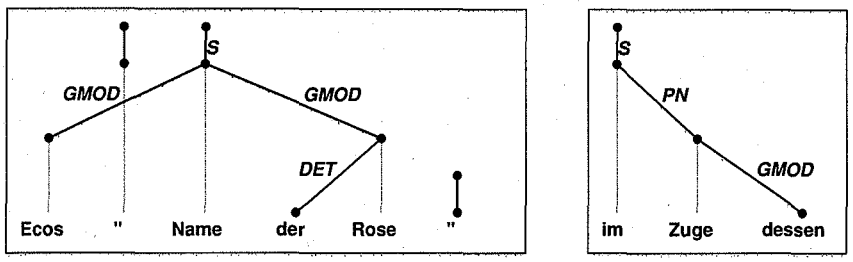
Bisweilen steht auch ein anderes Pronomen zusammen mit dem eigentlichen Objekt:



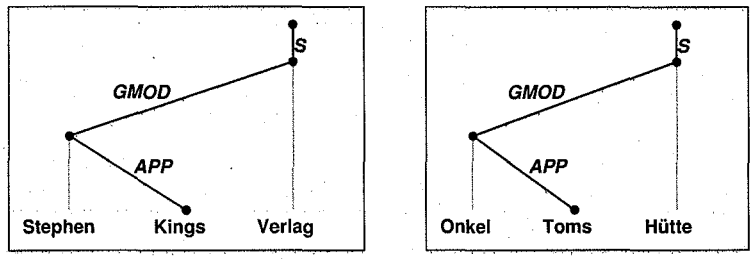
Möglicherweise ist dies auch ein expletiver Gebrauch. Derzeit erlaubt die Grammatik diese Struktur aber nicht.

Das Label GMOD

Dieses Label bezeichnet ein Genitivattribut. Sowohl das Attribut selbst als auch das modifizierte Wort ist ein Nomen oder ein substituierendes Pronomen.



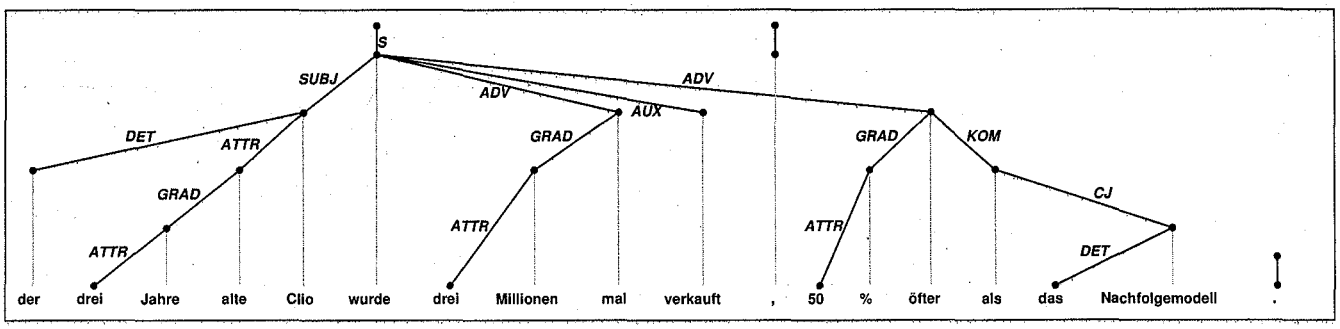
Das Genitivattribut ist immer ein Wort, das erkennbar im Genitiv steht; entweder das Wort selbst oder sein Determiner muß eine Genitivendung besitzen. Bei Eigennamen kann die Genitivendung auch am Ende des Namens stehen:



Das Label GRAD

Dieses Label bezeichnet eine im Akkusativ stehende NP, die als eine Art Maßangabe verwendet wird. Sie kann nur in ganz bestimmten Situationen auftreten:

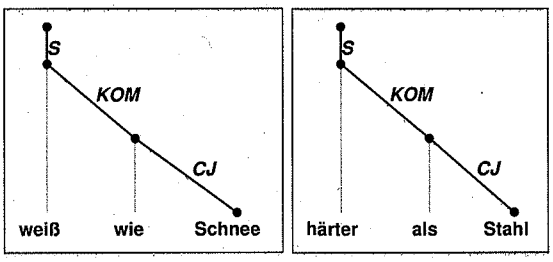
1. an einem Adjektiv, daß eine meßbare Eigenschaft ausdrückt ("hoch", "schnell"). Alle diese Adjektive tragen das Feature `measurement`.
2. an einem Komparativ ("mehr", "über")
3. als `CARD` mit dem Adverb "mal".
4. an einem Verb, das Bewegungsbedeutung hat: "Er versank einen Meter im Schlamm."



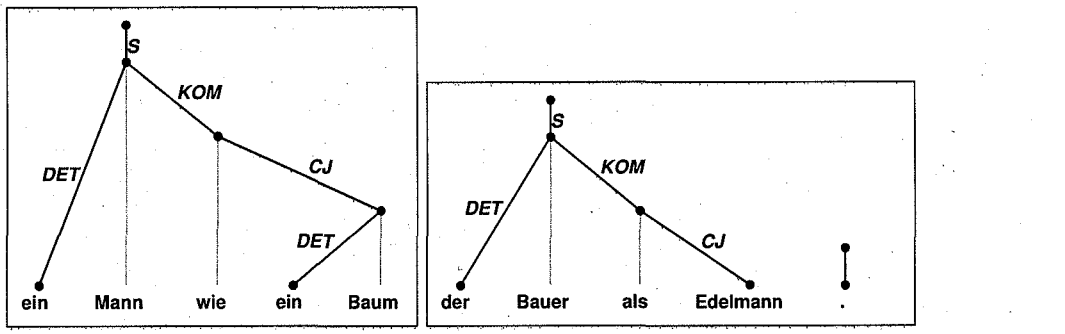
Die Maßangabe ist nur mit Zahlen und bestimmten Nomen möglich, nämlich Maßeinheiten (Jahr, Meter, Grad etc.).

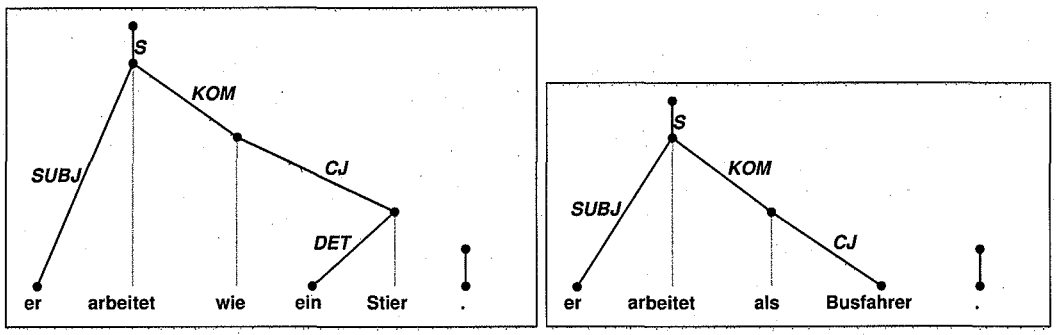
Das Label KOM

Mit KOM werden nur die Vergleichsworte "als" und "wie" untergeordnet. Sie können verschiedene Kategorien modifizieren. "als" modifiziert Adjektive und Pronomen im Komparativ, "wie" modifiziert Adjektive im Positiv.

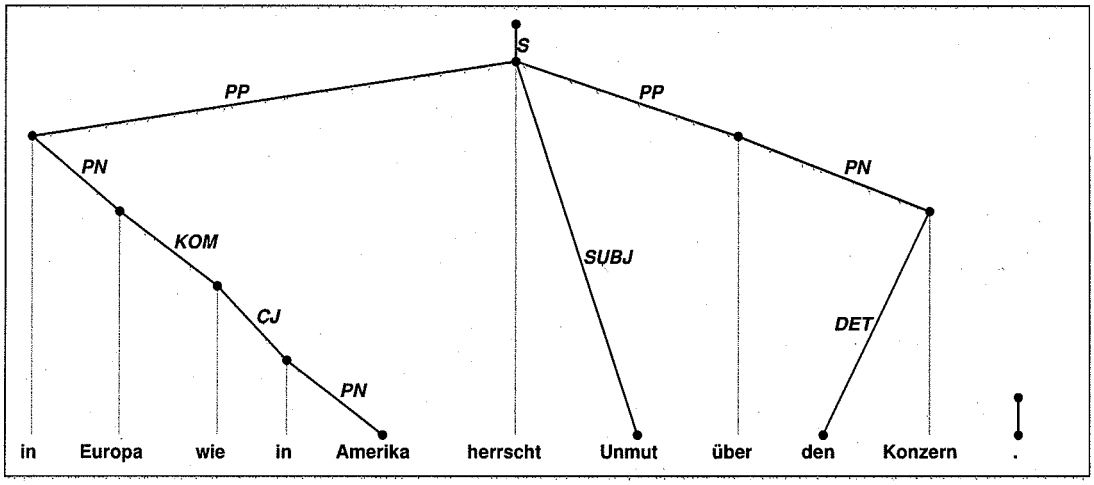


Beide Worte modifizieren Nomen und Verben:

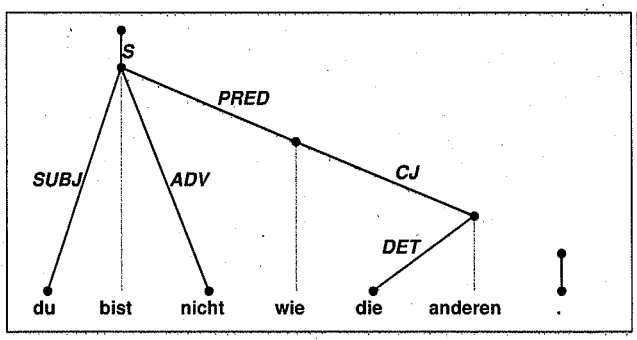


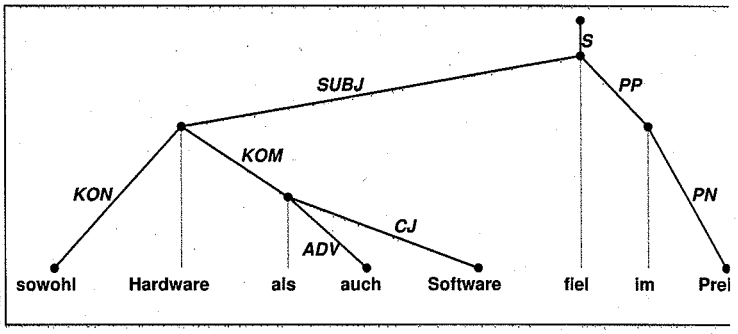


Das Wort "wie" kann auch Präpositionen modifizieren:



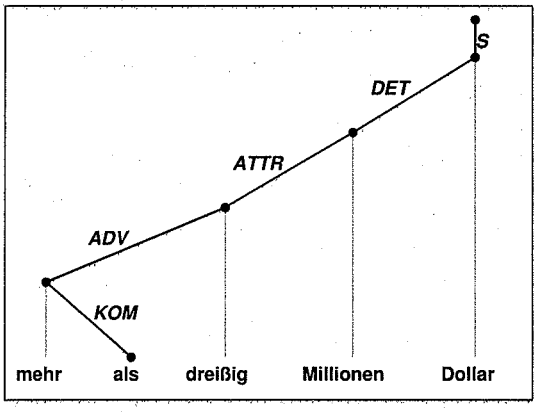
Beide Worte sind immer als KOM zu bezeichnen, wenn sie nicht Komplemente sind:





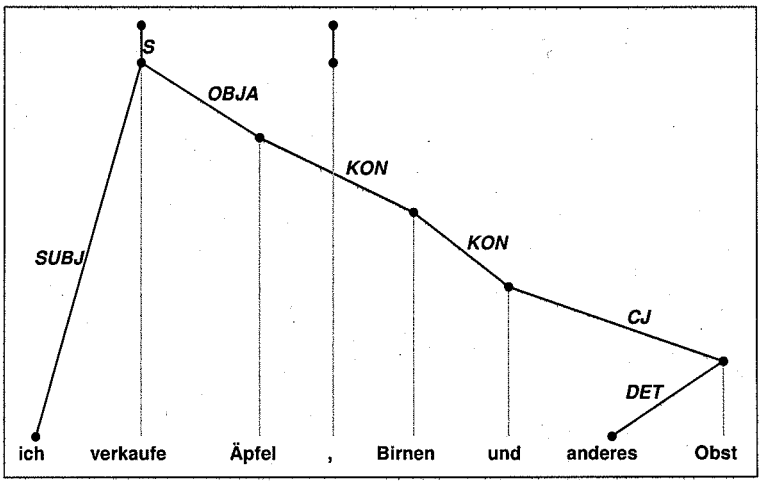
Nicht zu verwechseln mit den vergleichenden Konjunktionen "als" und "wie" sind die unterordnende Satzkonjunktion "als" ("als sieben Jahre vergangen waren") und das Fragewort "wie" ("wie geht es dir?").

Die Konstruktion 'mehr als' bzw. 'weniger als' wird stets als ein adverbialer Modifikator behandelt, also tief angebunden, so wie in der Formulierung 'über dreißig Millionen Dollar'.

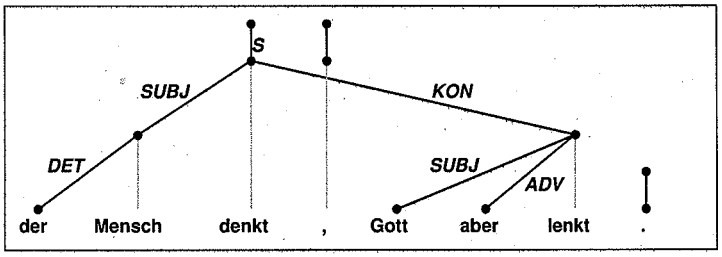


Das Label KON

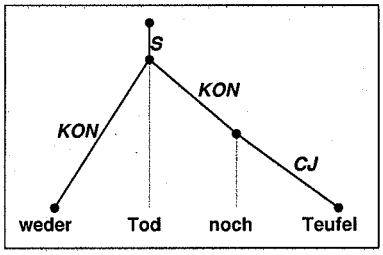
Beiordnungen werden als eine rechtsverzweigende Kette von Worten modelliert. Dabei trägt jedes Wort das Label KON, bis auf das letzte Wort, das unter der Konjunktion steht und mit CJ bezeichnet wird:



Einige Konjunktionen treten auch in adverbialer Form auf, d.h. nachgestellt. In diesem Fall sind sie als ADV zu bezeichnen, und die beigeordneten Worte modifizieren einander direkt als KON:



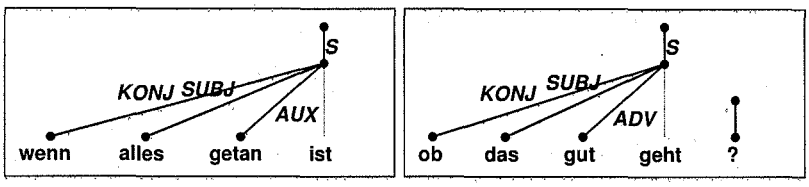
Einige Konjunktionen bestehen aus zwei Worten. Der linke Teil einer solchen Konjunktion wird als KON dem ersten Konjunkt untergeordnet:



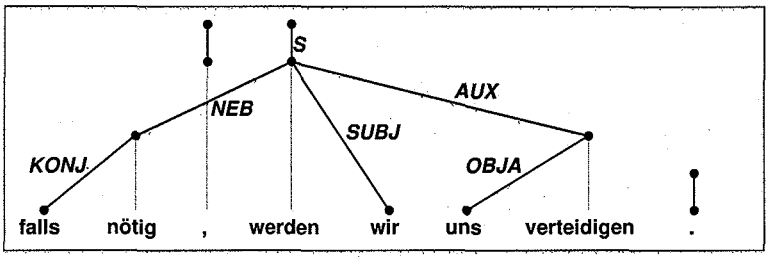
Das Label KONJ

Das Label KONJ verbindet eine unterordnende Konjunktion mit ihrem Bezugswort. Gewöhnlich ist das das Nebensatzverb.

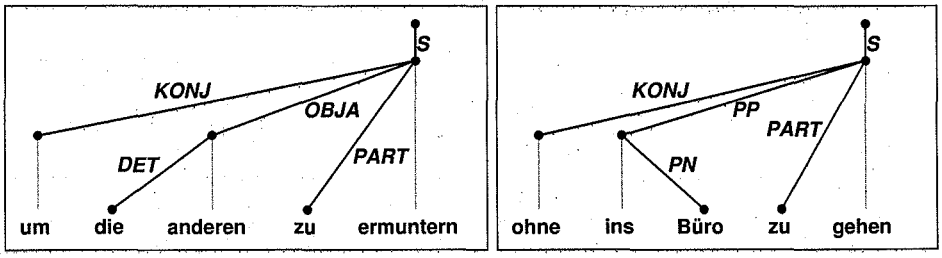
Satzkonjunktionen vom Typ KOUS modifizieren finite Verben:



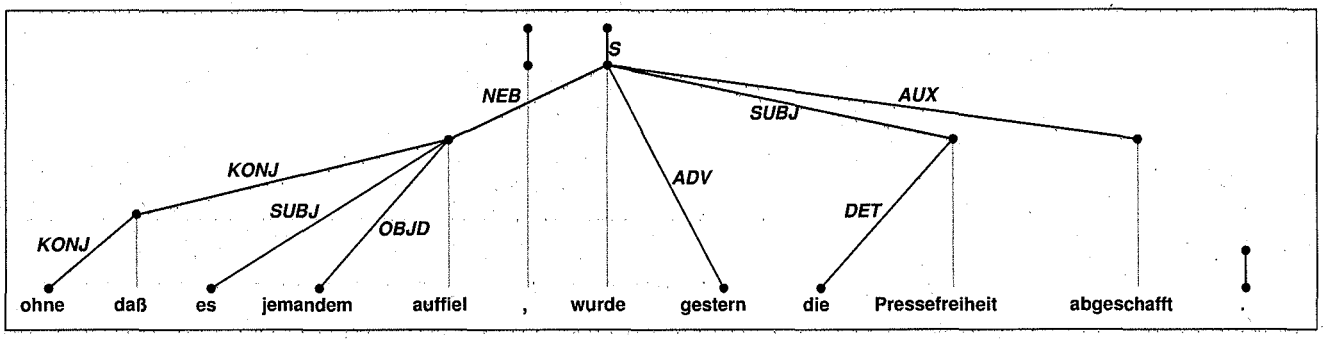
Bisweilen wird ein Nebensatz so weit verkürzt, daß nur noch ein Adjektiv verbleibt. In diesem Fall muß die Konjunktion das Adjektiv direkt modifizieren.



Konjunktionen vom Typ KOUI modifizieren erweiterte Infinitive:

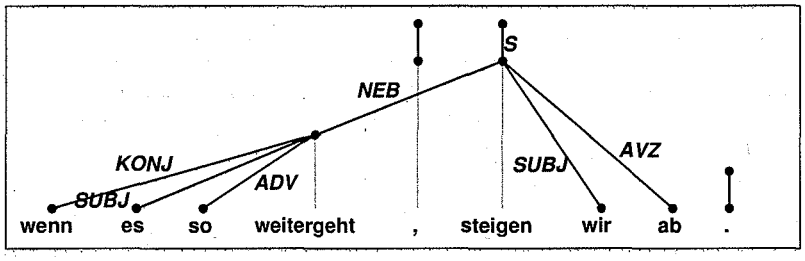


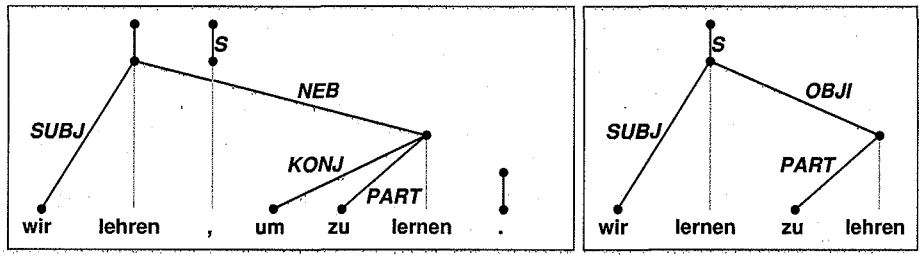
In den Konstruktionen "ohne daß" etc. soll die KOUI das "daß" modifizieren.



Das Label NEB

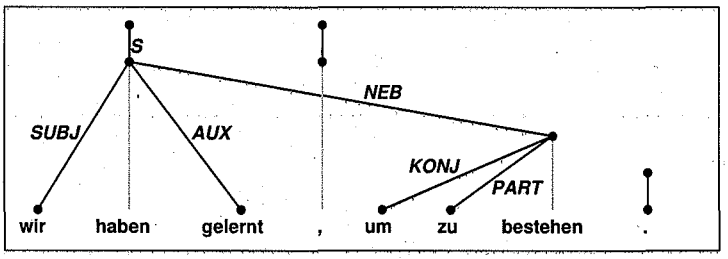
Das Label NEB verbindet das Verb eines Nebensatzes mit dem übergeordneten Wort.





Das untergeordnete Wort muß ein finites Verb oder ein erweiterter Infinitiv mit einer Konjunktion vom Typ `KOUI` sein. Erweiterte Infinitive ohne eine solche Konjunktion sind meistens Objektinfinitive (`OBJI`).

Wenn der Matrixsatz eine komplexe Verbgruppe besitzt, soll der Nebensatz das finite Verb modifizieren, nicht das Vollverb:

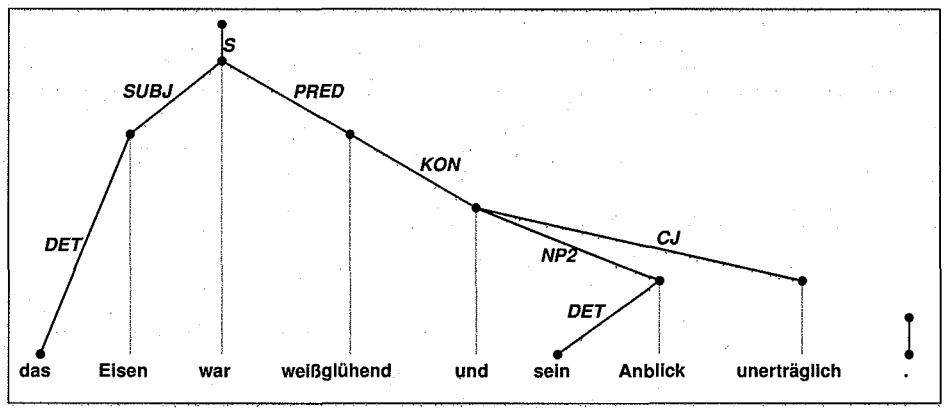


Das Label NP2

Das Label `NP2` dient dazu, ein überzähliges logisches Subjekt in einer elliptischen Koordination einzubinden, damit es nicht als Fragment behandelt werden muß. Typischerweise geschieht dies in Sätzen wie dem folgenden:

"Das Eisen war weißglühend und sein Anblick war unerträglich."

"Das Eisen war weißglühend und sein Anblick unerträglich."

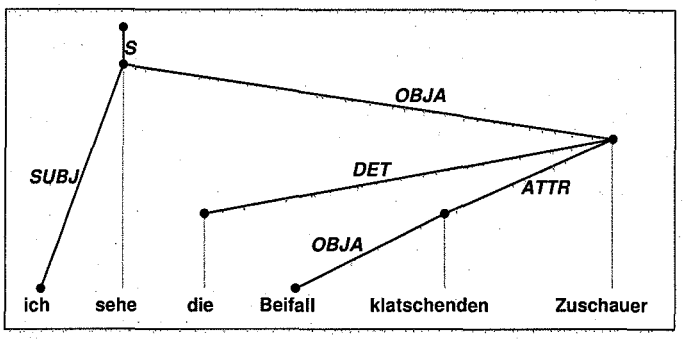


Diese Relation beschreibt allerdings nicht alle Fälle von Argument-Clustern richtig; beispielsweise läßt sich der folgende Satz nicht einmal durch `NP2` ohne Konflikt annotieren:

"Unsere Stärke war verbraucht, unsere Verbündeten geflohen, unser Wille gebrochen."

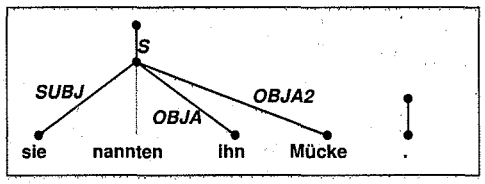
Das Label OBJA

Als Akkusativobjekt wird das zweite Komplement eines transitiven Verbs bezeichnet. Die Kante `OBJA` kann nur von Vollverben ausgehen (finit, infinit oder als Verbaladjektiv):



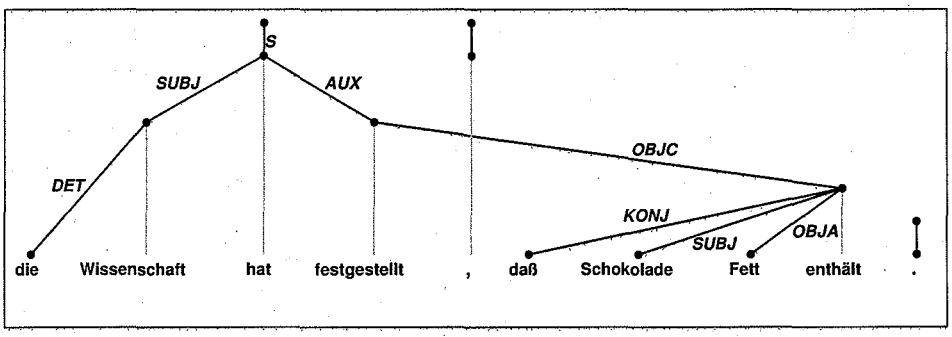
Das Label OBJA2

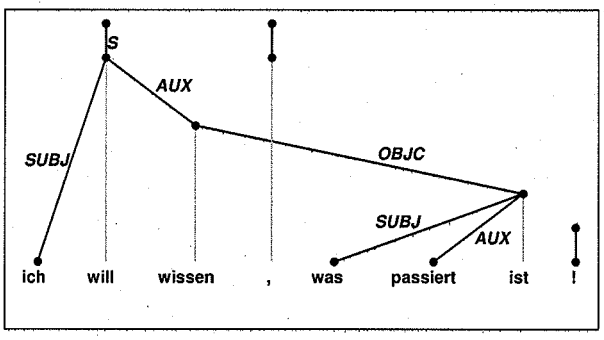
Dieses seltene Label kommt nur bei wenigen Verben vor, die zwei Akkusative verlangen (lehren, kosten, nennen).



Das Label OBJC

Ein Objektsatz ("object clause") ist ein finites Verb in Nebensatzstellung als Komplement eines anderen Verbs.

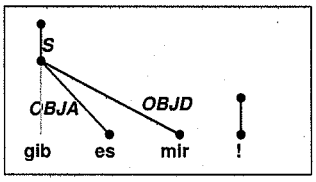


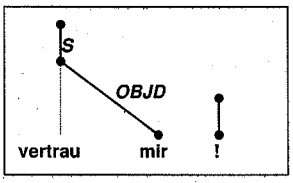


Solche Sätze treten nur mit 'daß' oder 'ob' oder Frageworten auf. Steht eine adverbiale Konjunktion am Verb, so handelt es sich um einen Nebensatz (NEB). Wenn der untergeordnete Satz in Hauptsatzstellung steht, ist er mit S zu bezeichnen.

Das Label OBJD

Das Dativobjekt kann entweder allein oder in Verbindung mit einem Akkusativobjekt auftreten. Was genau wo möglich ist, hängt vom Vollverb ab.

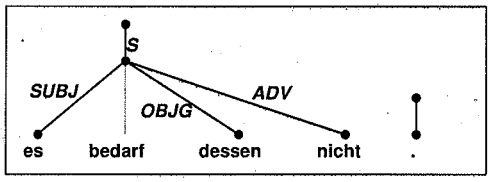




OBJD wird nur an Verben verwendet, die tatsächlich eine Dativvalenz besitzen. Andere Dativ-Unterordnungen sind meist ETH.

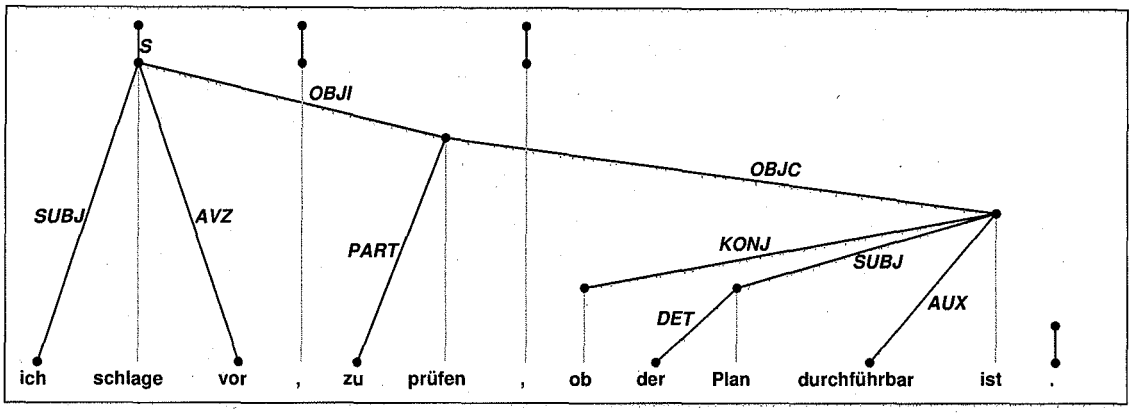
Das Label OBJG

Das Genitivobjekt ist das seltenste aller Objekte. Nur sehr wenige Verben verwenden es, und es ist immer obligatorisch.

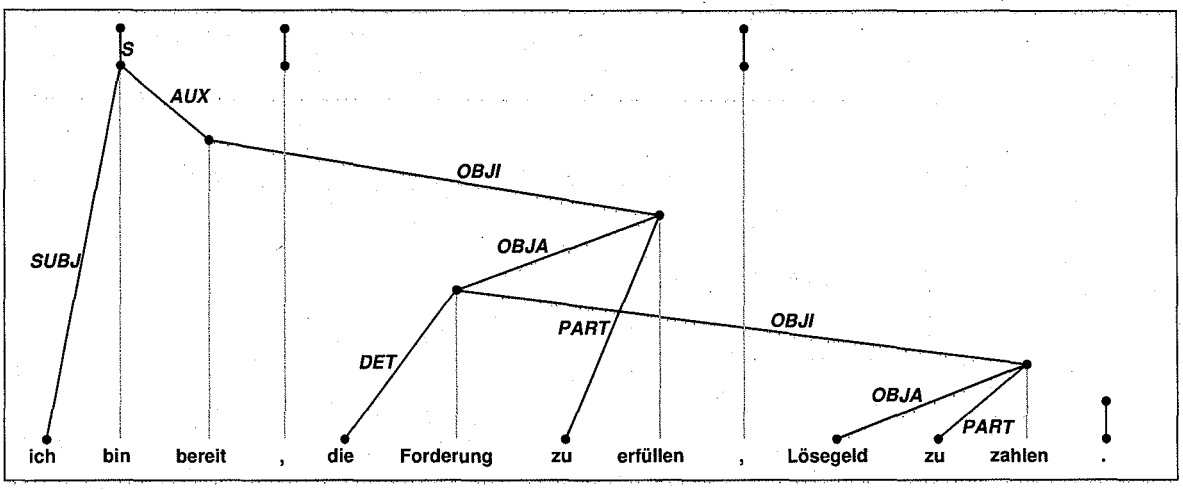


Das Label OBJI

Ein Objektinfinitiv ist ein Infinitiv als Komplement eines anderen Verbs. Nur bestimmte Verben können einen Objektinfinitiv tragen, und meistens kann er auch durch einen Akkusativ ersetzt werden.

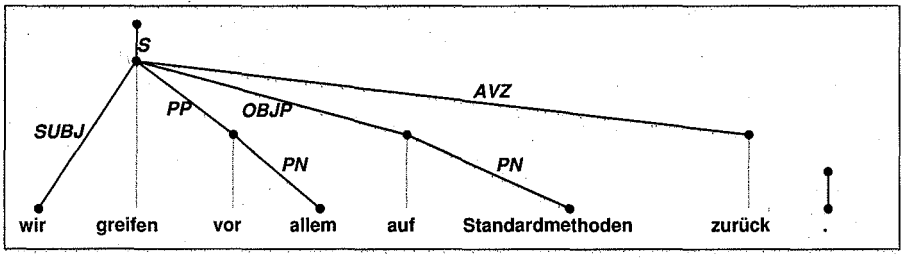


Bestimmte Nomen und Adjektive können ebenfalls Objektinfinitive nehmen. Meistens sind dies deverbale Wörter:



Das Label OBJP

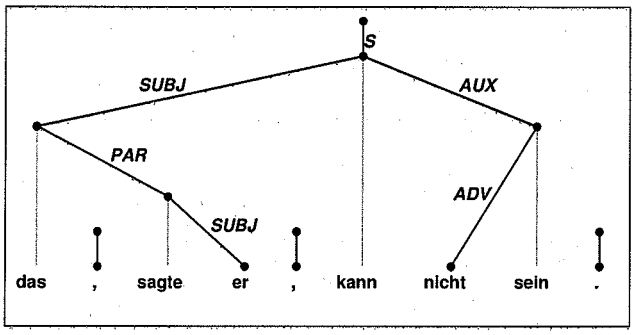
Als `OBJP` wird eine Präpositionalobjekt bezeichnet, also eine Präpositionalphrase, die das Komplement eines Verbs ist. Im Unterschied zu einer normalen `PP` darf sie nicht fortgelassen werden.



Das Label PAR

Dies Label bezeichnet *Parenthesen*, also Einschübe, von Matrixsätzen in den logisch untergeordneten Objektsatz. Immer wenn die Unterordnung eines Objektsatzes einen nichtprojekti-

ven Baum erzeugen würde, wird statt dessen der Matrixsatz dem Objektsatz untergeordnet. Das Label ist dann PAR, und der Regent ist das letzte Wort vor der Parenthese.

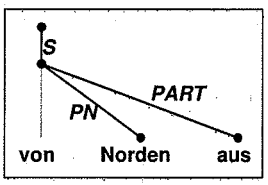


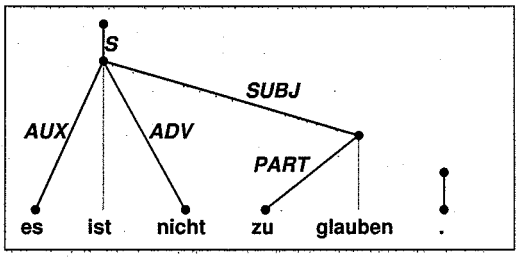
Es gibt verschiedene andere Konstruktionen, die manchmal Parenthese genannt werden; diese werden **nicht** als PAR bezeichnet, sondern so wie immer. Beispielsweise ist eine Präpositionalphrase PP und nicht PAR, selbst dann, wenn sie nicht in die normale Satzstruktur paßt, etwa im Vor-Vorfeld.

Das Label PART

Dieses Label bezeichnet die Unterordnungen von mehreren sehr stark eingeschränkten Partikeln:

1. eine Zirkumposition (APZR) modifiziert eine Präposition
2. das Wort "zu" (PTKZU) modifiziert einen Infinitiv

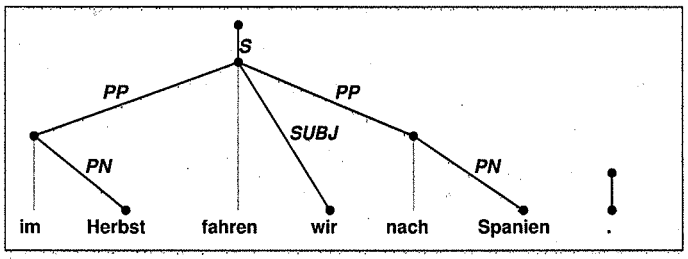




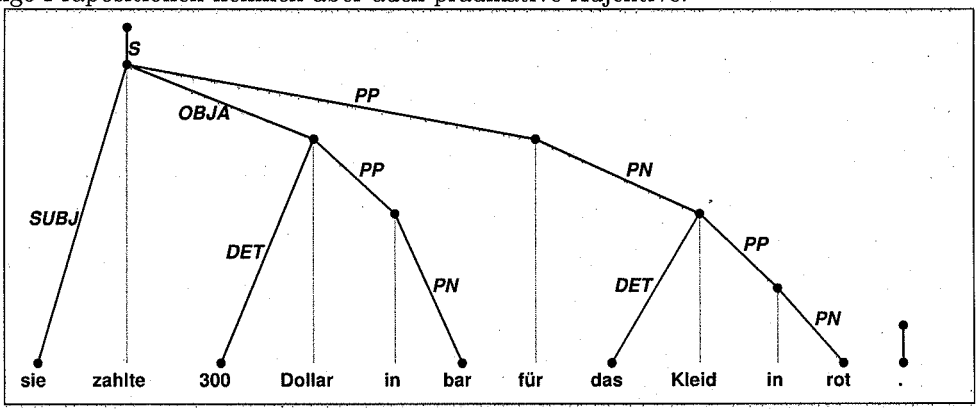
(Im ersten Fall ist die Partikel genau genommen ein Komplement und kein Modifikator; aber im Deutschen gibt es nur solche Zirkumpositionen, die auch als einfache Präpositionen dienen können, daher wird die Partikel immer als optional behandelt.)

Das Label PN

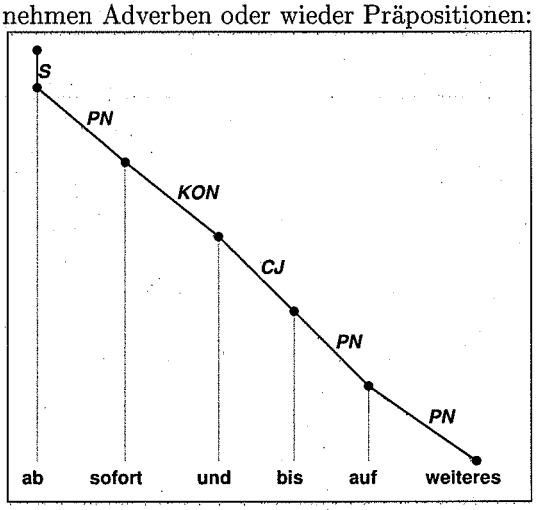
Dieses Label bezeichnet das Komplement einer Präposition. Gewöhnlich sind das Nomen und Pronomen:



Einige Präpositionen nehmen aber auch prädikative Adjektive:

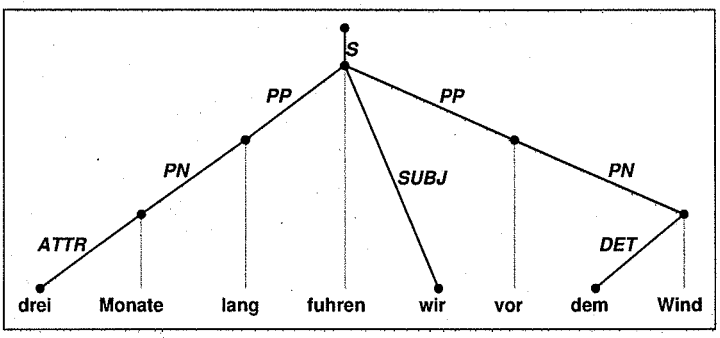


Andere Präpositionen nehmen Adverben oder wieder Präpositionen:



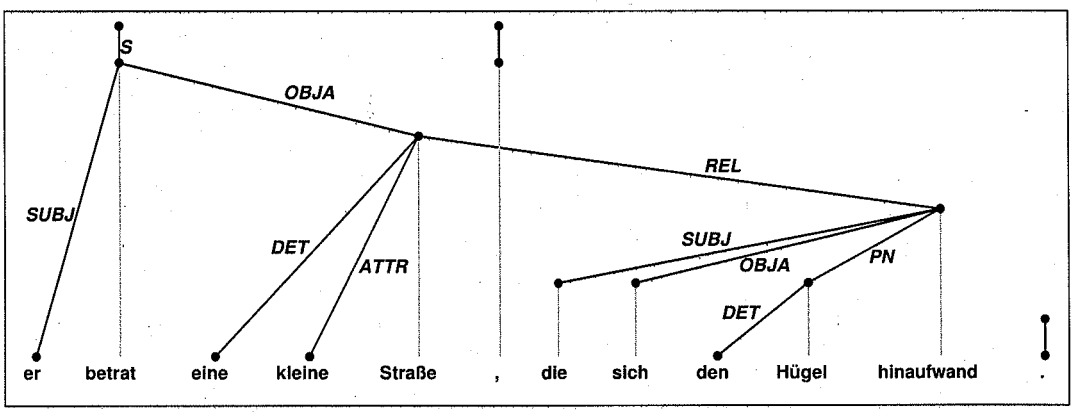
Hierbei handelt es sich fast immer um feststehende Fügungen.

Die wenigen Postpositionen, die das Deutsche aufzuweisen hat, nehmen ebenfalls Argumente mit dem Label PN:

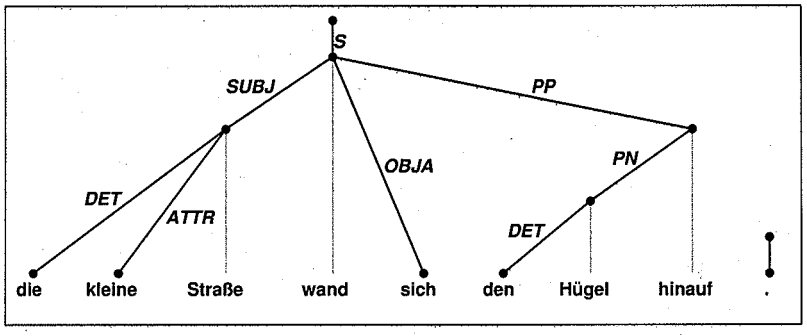


Die Vergleichsworte "wie" und "als" gelten nicht als Präpositionen; deshalb werden ihre Komplemente nicht mit PN bezeichnet, sondern mit CJ.

Einige Präpositionen und Postpositionen werden gelegentlich ohne Worttrennung an ein Vollverb angehängt. Es handelt sich hierbei um die Adpositionen 'herunter', 'herauf', 'herab', 'hinunter', 'hinauf', hinab' und 'entlang'. Obwohl sie sich also wie abtrennbare Präfixe verhalten, sind es doch immer noch Adpositionen, was daran erkenntlich ist, daß sie noch immer eine Valenz für eine weitere NP eröffnen. Bei diesen seltenen Fällen (und nur in der Nebensatzstellung) kann daher auch ein Vollverb durch PN modifiziert werden:



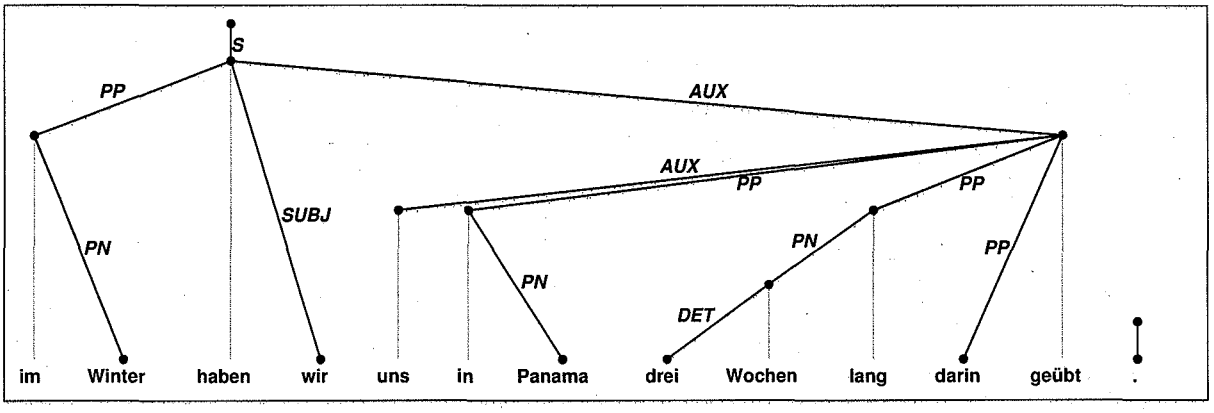
In der Hauptsatzstellung ist die Adposition als APPO anzusehen und bildet daher eine normale PP:



Das Label PP

Mit diesem Label werden Präpositionalphrasen untergeordnet. Außer normalen Präpositionen (APPR) können verschmolzene Präpositionen (APPRART), Postpositionen (APPO) und

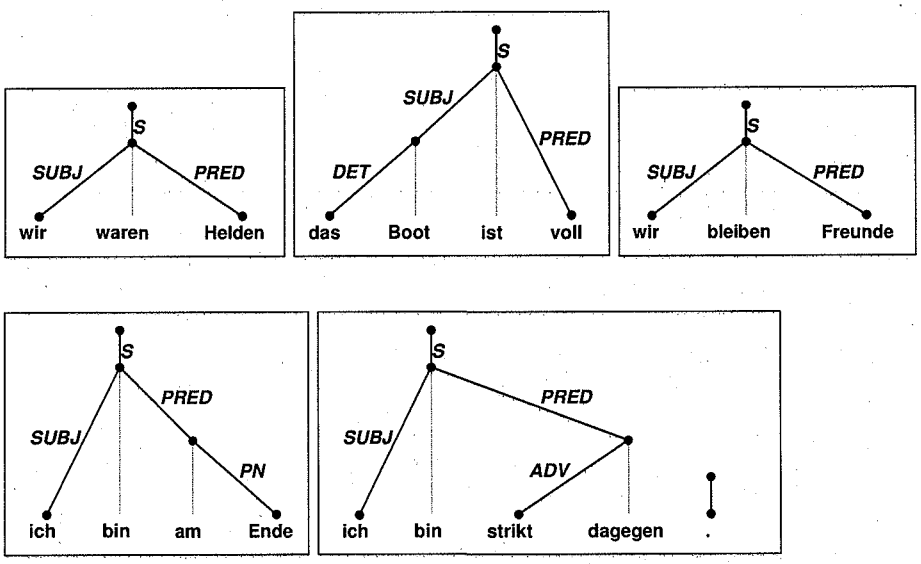
Pronominaladverben (PROAV) eine Präpositionalphrase einleiten.

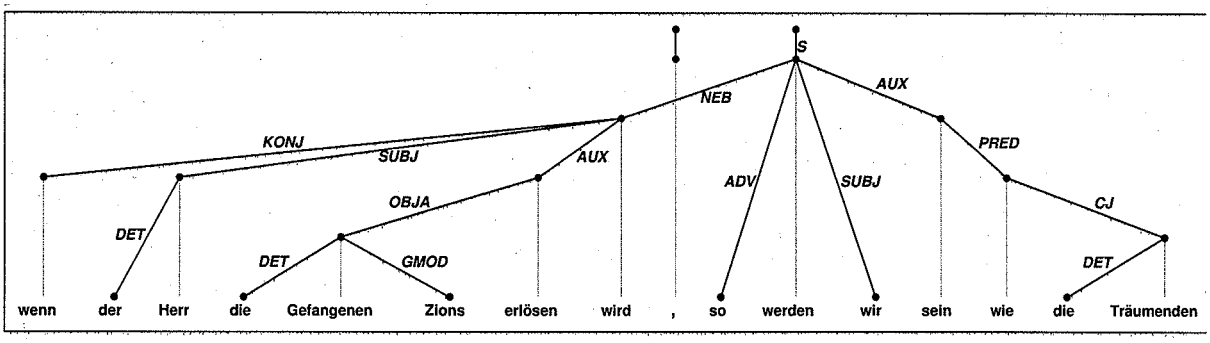


Präpositionalphrasen modifizieren vorwiegend Verben, weniger oft Nominalphrasen und Adjektive. Der Bezug ist oft nicht eindeutig zu klären.

Das Label PRED

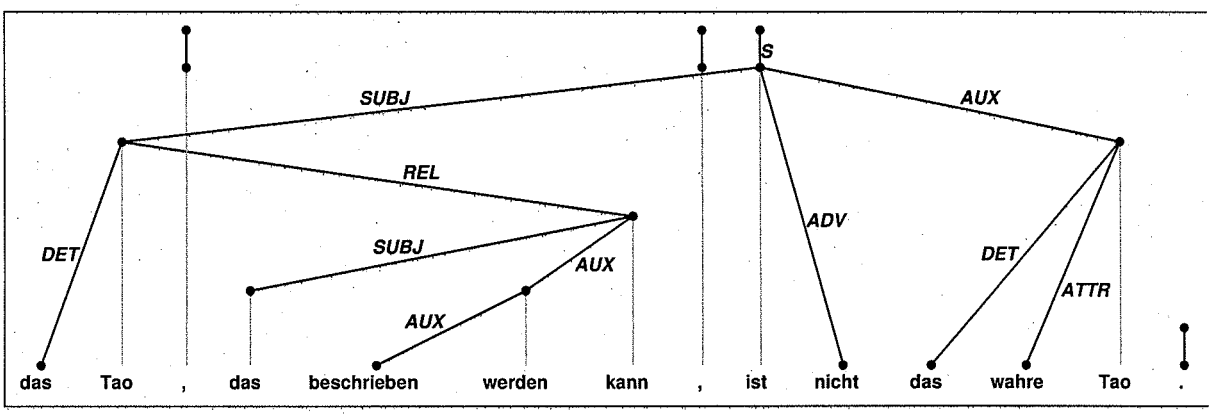
Dies Label bezeichnet prädikative Ergänzungen, typischerweise am Verb 'sein' oder ähnlichen Verben wie 'werden', 'wirken' oder 'scheinen'. Meistens ist das Prädikat ein prädikatives Adjektiv oder eine NP, es kann aber, insbesondere am Verb 'sein', auch Adverb oder Präposition sein:

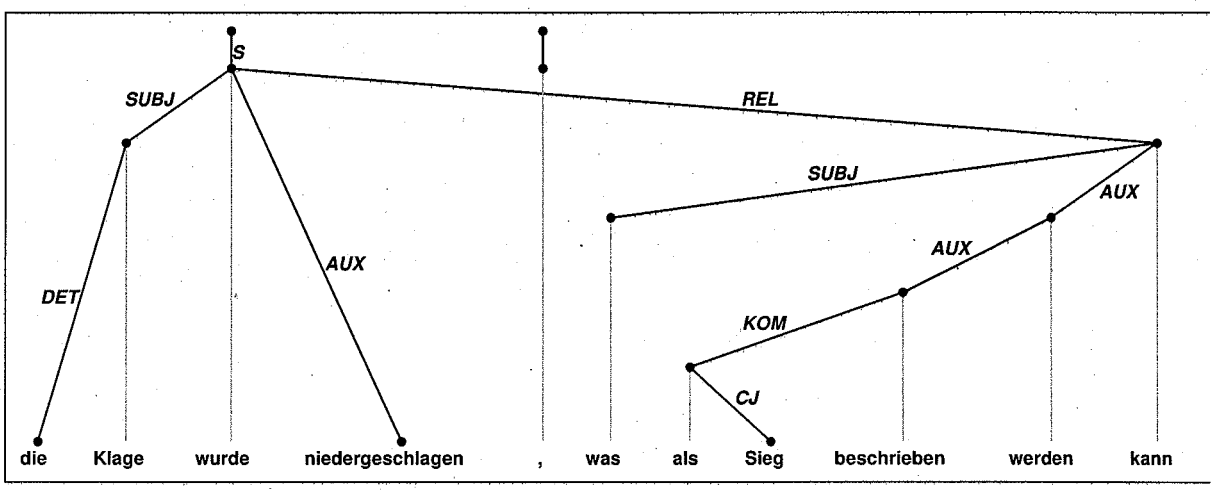




Das Label REL

Dieses Label bezeichnet die Beziehung zwischen dem Verb eines Relativsatzes und dem übergeordneten Wort. Das untergeordnete Wort ist immer ein finites Verb. Das übergeordnete Wort kann dagegen ein Verb oder ein Hauptwort sein.

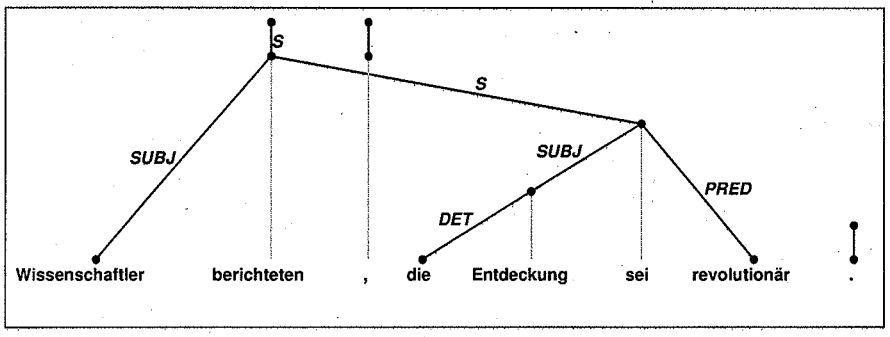




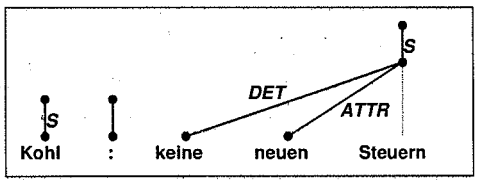
Ein Relativsatz ist immer ein Nebensatz, der mit einem Relativpronomen ("der") oder einem Fragewort ("womit") eingeleitet wird.

Das Label S

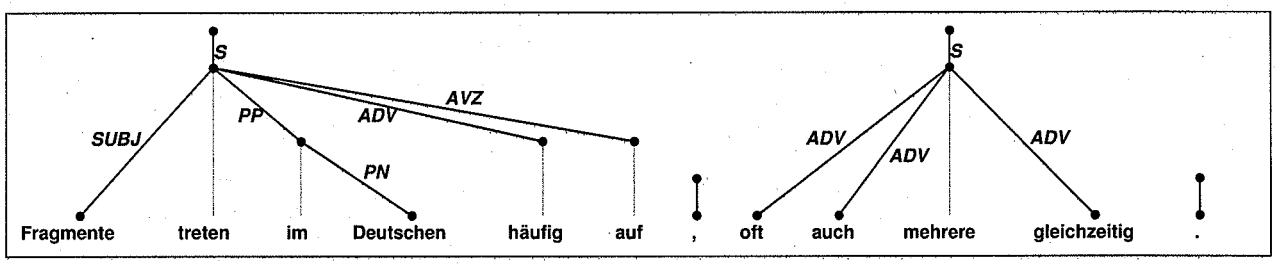
Das Label S bezeichnet das Wurzelwort eines vollständigen Satzes.



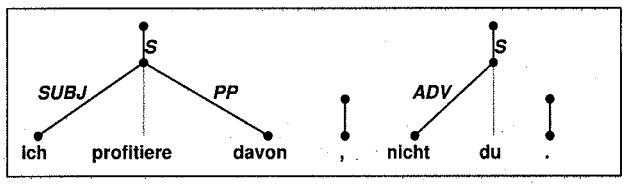
S wird auch für andere nicht untergeordnete Worte[1] verwendet, also für das dominierende Wort eines Fragmentes. Ein Fragment ist eine Konstruktion, die nicht in eine Satzstruktur eingeordnet werden kann, ohne harte Bedingungen zu verletzen. Typischerweise tauchen Fragmente in Überschriften auf:



Oftmals werden Fragmente mit Komma abgetrennt und durch Adverben markiert:

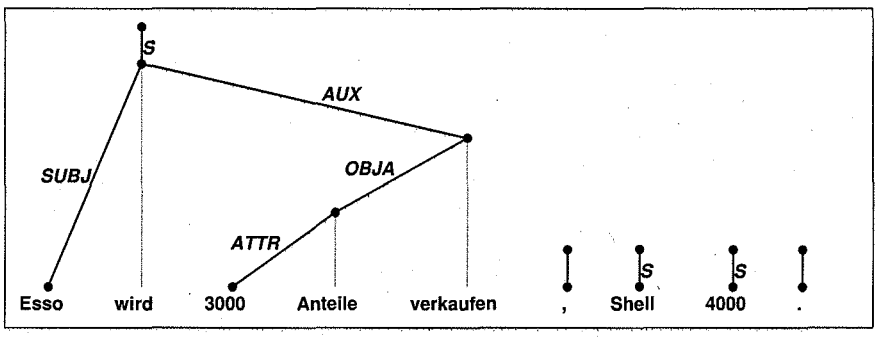


Eine weitere Quelle von Fragmenten sind elliptische Koordinationen. Oft wird etwa vor "nicht" das "und" fortgelassen:



In anderen Fällen werden mehrere gleichlautende Worte aus einer Koordination weggelassen, weil sie sonst eine Wiederholung zur Folge hätten:

[1] Satzzeichen gelten nicht als Worte; sie tragen daher auch kein Label. (Genaugenommen tragen sie den leeren String als Label.)

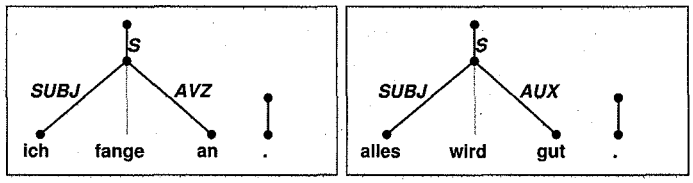


Solche Konstruktionen sind in einer rein deklarativen Dependenzgrammatik schlecht zu modellieren. Meistens müssen die verbleibenden Wörter als Fragmente angesehen werden.

Wenn ein ganzer Nebensatz isoliert auftritt, bleibt sein Label NEB, nicht S; dasselbe gilt für REL.

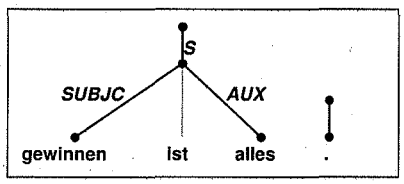
Das Label SUBJ

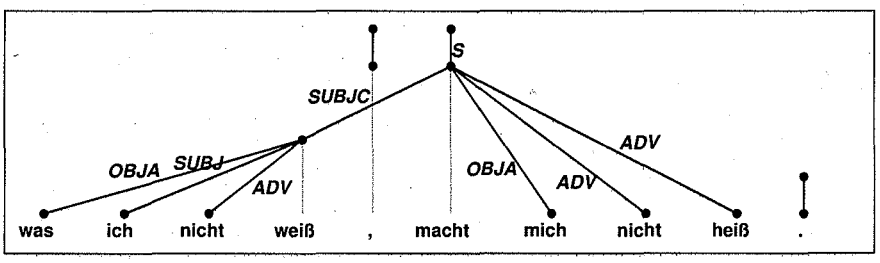
Als SUBJ wird das Subjekt eines finiten Verbs bezeichnet. Das ist immer irgendeine Art von NP:

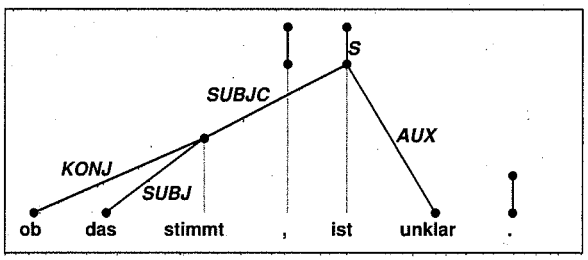


Das Label SUBJC

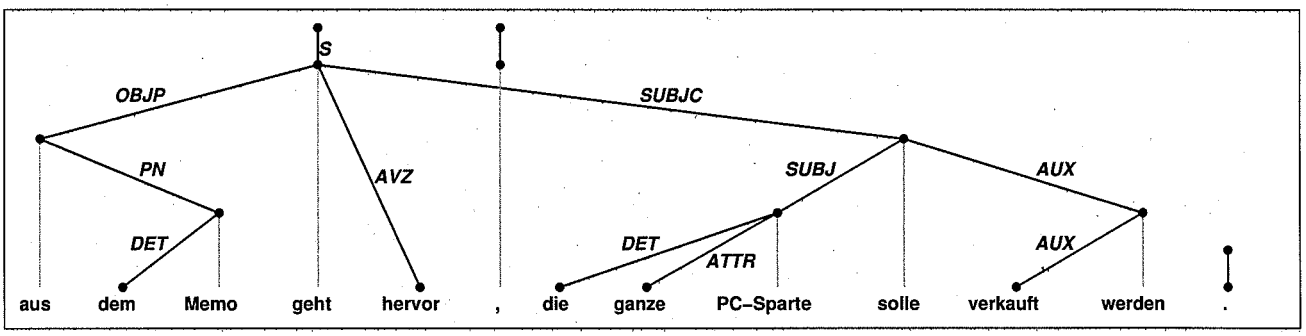
Es gibt jedoch auch ganze Subjektsätze, d.h. infinitive oder finite Verben, die die Subjektrolle ausfüllen. Diese werden als SUBJC bezeichnet.







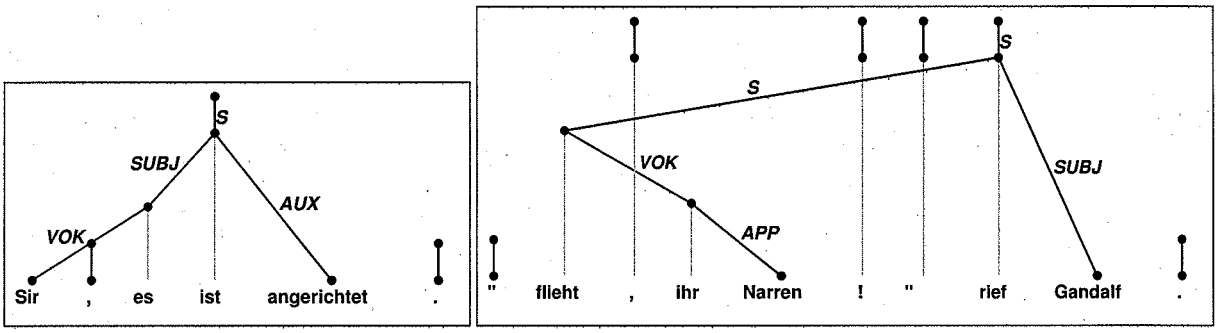
Ein finites Verb kann gewöhnlich nur dann die Subjektrolle ausfüllen, wenn es mit "dass", "ob" oder einem Fragewort verwendet wird. In seltenen Fällen kann ein ganz normaler (konjunktionsloser) Hauptsatz Subjektsatz sein. Das ist jedoch nur mit ganz bestimmten Verben möglich:



Alle diese Verben tragen das Feature `nimmt_Subjektsatz`.

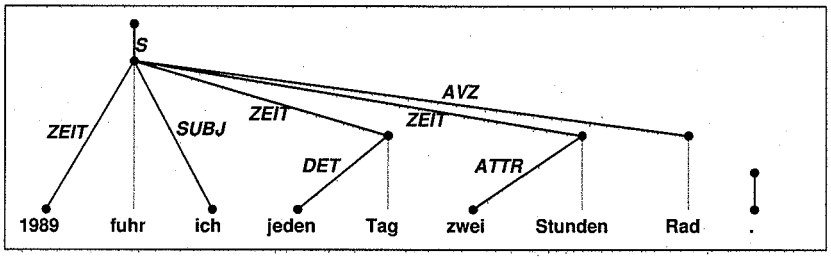
Das Label VOK

Das Label `VOK` verbindet Anreden mit dem Satz, zu dem sie gehören. Es wird jeweils am nächstliegenden Wort untergeordnet, egal welcher Kategorie dieses angehört.



Das Label ZEIT

Dieses Label bezeichnet konjunktionslose Zeitangaben. Solche Formulierungen stehen immer im Akkusativ und sind auf sehr wenige Nomen und Zahlen beschränkt.

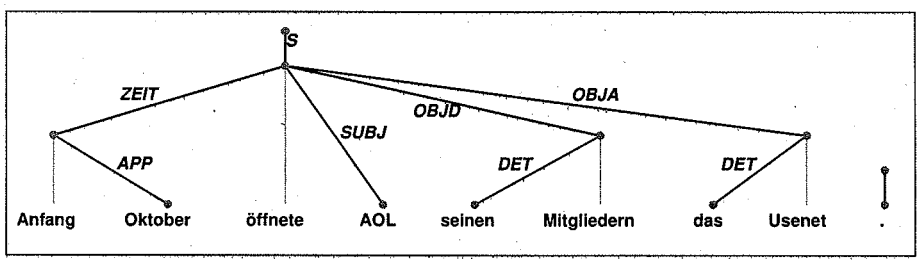


Nur folgende Worte können als Zeitangabe verwendet werden:

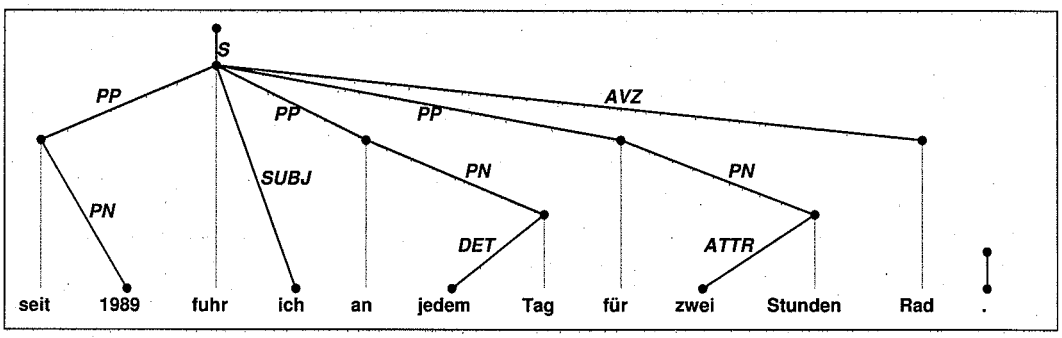
- Kardinalzahlen, wenn sie als Jahreszahlen verwendet werden
- "Anfang", "Mitte" und "Ende" in Verbindung mit anderen Zeitangaben
- Zeiteinheiten ("Tag", "Woche" etc.) und benannte Zeiten ("Montag", "Januar" etc.)

All diese Worte zeichnen sich dadurch aus, daß sie bestimmten semantischen Klassen angehören (feature `sort`).

Wenn komplexe NP als Zeitangaben auftreten, ist nur die oberste Kante `ZEIT` und alle anderen normal `APP`.



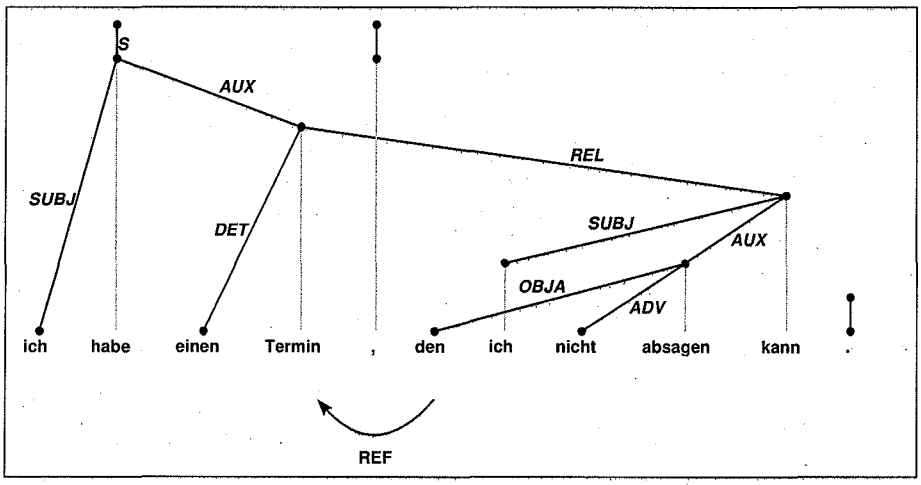
Wenn eine solche Zeitangabe mit einer Präposition verwendet wird, ist sie nicht als `ZEIT` zu kennzeichnen, sondern als `PN`:



1.1.2 Die Referenz-Ebene

Diese Analyseebene stellt Referenzbeziehungen zwischen Pronomen und ihren Antezedenten dar. Derzeit werden nur Relativpronomen berücksichtigt (obwohl prinzipiell auch Personalpronomen oder auch ganze Nominalphrasen auf andere Konstituenten referieren können).

Die einzige mögliche Kante auf dieser Ebene ist die vom Relativpronomen zu seinem Antezedent. Meistens ist dies das Nomen, das vom Relativsatz modifiziert wird:



1.2 Annotationsrichtlinien

Im Prinzip ist die gewünschte Struktur für jede Äußerung durch die Constraints der Grammatik definiert. Da die Constraints aber Fehler enthalten können und außerdem weiterentwickelt werden, ist hier ausdrücklich beschrieben, wie in Zweifelsfällen annotiert werden soll.

1.2.1 Was ist ein Satz?

Grundsätzlich kann CDG nur einzelne Sätze analysieren und nicht einen ganzen Text. Gewöhnlich wird ein Satz durch den Satzpunkt oder durch ähnliche Satzzeichen wie ! oder ? beendet. Dabei kann ein 'Satz' für CDG durchaus auch mehrere Hauptsätze enthalten, wenn diese beigeordnet, untergeordnet oder gar ineinander verschränkt sind:

- Pandesic bediene nur ein geringes Marktsegment für kleine Firmen, sagte der SAP-Sprecher Michael Pfister.
- Tagesschau-Moderatoren geben einen Einblick in die Nachrichtensendung, Günther Jauch spielt mit Gästen "Wer wird Millionär?", die Teenie-Gruppe "No Angels" tritt auf und Verona Feldbusch spricht mit ihrem Cyber-Double.
- Bis zum Jahresende 2002, prognostiziert Roland Berger, werden die am Neuen Markt gelisteten Unternehmen 200.000 Mitarbeiter beschäftigen.

Alle diese Beispiele sollen als jeweils ein Satz annotiert werden, wobei die verschiedenen Hauptsätze mit Labeln wie `S`, `KON` oder `PAR` zu verbinden sind.

In manchen Fällen sind mehrere Hauptsätze derart ineinander verschränkt, daß sie nur durch Umordnung voneinander getrennt werden könnten. In solchen Fällen ist ebenfalls ein

Satz anzunehmen; können die unabhängigen Bestandteile aus strukturellen Gründen nicht in einen Syntaxbaum zusammengefaßt werden, so müssen mehrere Satzwurzeln annotiert werden:

- Der Gerichtshof der Europäischen Union ist nach Maßgabe der folgenden Bestimmungen zuständig in Streitsachen über a) die Erfüllung der Verpflichtungen der Mitgliedstaaten aus der Satzung der Europäischen Investitionsbank. Der Verwaltungsrat der Bank besitzt hierbei die der Kommission in Artikel III-360 übertragenen Befugnisse; b) die Beschlüsse des Rates der Gouverneure der Europäischen Investitionsbank.

1.2.2 Was ist ein Wort?

Normalerweise werden Worte durch Leerzeichen oder Satzzeichen voneinander abgegrenzt, wobei jedes Satzzeichen als ein Wort gilt.

- "Was wollt ihr?" schrie er. (neun Worte)

Es gibt jedoch einige Ausnahmen. Mehrteilige Namen gelten bald als mehrere Worte, bald als ein Wort; wird aber ein Adjektiv aus einem solchen Namen gebildet, so gilt es immer als ein Wort.

- Bad Nauheim (zwei Worte)
- der Bad Nauheimer Kulturverein (drei Worte)

Apostrophe, die Auslassungen oder Genitiv anzeigen, gehören zum betreffenden Wort. Dieses selbst wird aber einzeln gezählt, auch wenn es mit dem vorigen zusammengeschrieben wird.

- Leibniz' Philosophie des Unabwendbaren (vier Worte)
- Das war's. (vier Worte)

Sind Anführungszeichen teilweise in Zusammensetzungen verschmolzen, so gehören beide Anführungszeichen zum Wort dazu.

- kurze "Jingle"-Einspielungen (zwei Worte)

Das entsprechende gilt für modische Klammerungen.

- eine/n Schüler/in (zwei Worte)
- man(n) amüsiert sich (drei Worte)
- der mühsame 2:1(1:1)-Erfolg (drei Worte)

Leider sind solche Zusammensetzungen nicht auf kurze Phrasen beschränkt:

- Der Konzern fährt eine hochprofitable "wir kümmern uns nicht um Standards"-Strategie.

Für solche Konstruktionen gibt es keine angemessene Modellierung. Entweder muß man ein einziges bizarres Wort annehmen, was die interne Struktur der Beschwerde völlig ignorieren würde, oder sie normal annotieren, wonach sie aber nicht mehr in den äußeren Satz untergeordnet werden kann. Die richtige Lösung (die Beschwerde intern als Hauptsatz mit Label S zu annotieren und nach außen hin als ein TRUNC zu klassifizieren) setzte eine Mehr-Ebenen-Struktur voraus, die CDG nicht erlaubt.

Satzzeichen, die zu Eigennamen gehören, gehören zum betreffenden Wort.

- Wham!, die Gruppe mit dem Ausrufezeichen (sieben Worte)

Komposita mit Bindestrichen gelten als ein Wort. Wenn ein solches Wort am Zeilenende getrennt wurde, ist es wieder zusammenzusetzen.

- Bayern, Baden-Württemberg und Mecklenburg-Vorpommern (vier Worte)

Arabische Zahlen gelten als ein Wort, auch wenn sie mit Kommas oder Punkten unterteilt sind.

- 1,000,000 Millionen Dollar (drei Worte)

Manche Zeitungen verwenden gar Leerzeichen anstelle von Punkten; in diesem Fall sollte die Zahl wieder zusammengesetzt werden.

1.2.3 Welche syntaktische Kategorie?

Die Einteilung von Worten in syntaktische Kategorien folgt den Richtlinien des STTS (vgl. 'Guidelines für das Tagging deutsche Textcorpora mit STTS'). Wo dieses Dokument unvollständig oder widersprüchlich ist, gelten folgende Richtlinien. Niemals sollte eine Kategorie nur deshalb gewählt werden, weil der Tagger sie vorhersagt!

ADJA oder NN?

Praktisch alle Adjektive können substantiviert werden. Gemäß STTS werden sie als ADJA eingeordnet, wenn sie klein geschrieben sind, sonst als NN.

- Die kleinen/ADJA hängt man und die großen/ADJA läßt man laufen.
- Die Roten/NN haben nichts Großes/NN bewirkt.

Attributiv gebrauchte Adjektive sind immer ADJA, ob sie groß oder klein geschrieben sind.

- die hohe/ADJA Steuer
- der Hohe/ADJA Rat

ADJD oder APPR?

Einzelne Adjektive tragen Nominalargumente, als wenn sie Präpositionen wären:

- Der Bauernhof liegt südlich/ADJD der Stadt.
- Dieses Verhalten ist eines Prinzen unwürdig/ADJD.
- Wir waren das Kämpfen müde/ADJD.

Laut STTS ist aber nur eine ganz bestimmte, nicht sonderlich logisch begründbare Menge von Worten APPR. Im Einzelfall muß also in der Wortliste dort nachgesehen werden, ob es sich um APPR oder um ADJD handelt. Je nachdem sind die Kanten dann mit PP und PN oder mit ADV und OBJG (oder OBJA oder OBJD) zu bezeichnen.

- Der Bauernhof liegt oberhalb/APPR der Stadt.
- Dieses Verhalten ist gemäß/APPR seiner hohen Herkunft.
- Diese Verordnung gilt vorbehaltlich/APPR der Verkündung im Bundestag.

ADJD oder VVPP?

Partizipien werden laut STTS als ADJD eingeordnet, wenn sie adverbial gebraucht sind, sonst als VVPP.

- Der Feind wurde gestellt/VVPP.
- Der Feind ist gestellt/VVPP.
- Die Szene wirkt gestellt/ADJD.
- Die Szene ist gestellt/ADJD.

Wann liegt nun ein adverbialer Gebrauch vor? Wenn die Bedeutung offensichtlich passivisch ist, ist auch bei 'sein' ein Partizip anzunehmen:

- Der Tisch ist verrückt/VVPP. (= Man hat den Tisch verrückt)
- Die Tante ist verrückt/ADJD. ($\neq$ Man hat die Tante verrückt)

In vielen Fällen ist die Unterscheidung nicht klar. Das STTS gibt eine lange Liste mit Partizipien an, die fallweise ADJD sein können; diese ist weder vollständig noch ausschließlich. Keine Angabe macht es zu folgender Konstruktion:

- Das waren kleine Stifte, gedrechselt/VVPP aus Buchenholz.

In solchen Sätzen, die aus Relativsätzen verkürzt wurden, wählen wir stets VVPP.

ADV oder APPR?

Einige quantifizierende Adverben haben Homonyme unter den Präpositionen. Das sind die Worte 'bis', 'gegen', 'unter', 'über', und 'zwischen'. Die Adverb-Lesart tritt nur auf, wenn das Wort direkt vor einer Zahl steht. Auch dann kann aber immer noch die Präpositions-Lesart eintreten. Daher muß die Satzaussage geprüft werden: kann der Satz paraphrasiert werden durch 'nicht mehr als', 'ungefähr', 'mehr als', 'weniger als' oder 'von...bis', so liegt die Adverb-Lesart vor. Also:

- Hamburg besitzt über/ADV tausend Brücken
- Hamburg verfügt über/APPR tausend Brücken.

In manchen Fällen widersprechen sich Morphologie und Semantik:

- In diese Kategorie fällt nur jeder neunte Deutsche (11 Prozent der Gesamtbevölkerung über 14 Jahren).

Hier ist die Bedeutung von 'über' eindeutig 'mehr als'; dennoch ist 'Jahren' im Dativ gebraucht. Auch hier ist dennoch ADV zu wählen.

ADV oder KON?

Gemäß STTS sind die Konjunktionen 'aber', 'doch', 'denn' und 'jedoch' als KON zu bezeichnen, wenn sie zwischen zwei Konstituenten stehen, jedoch als ADV, wenn sie in die zweite eingeschoben sind. Diese Unterscheidung wird befolgt, obwohl sich durch die Umstellung keinerlei Sinnänderung ergibt.

- Wir haben den Krieg gewonnen, aber/KON den Frieden verloren.
- Wir haben den Krieg gewonnen, den Frieden aber/ADV verloren.

Die zweite Konstituente selbst wird aber in beiden Fällen als 'KON' untergeordnet.

ADV oder PTKVZ?

Die Unterscheidung zwischen freien Adverben und abgetrennten Verbzusätzen ist schwer zu treffen, weil diese aus jenen entstanden sind. Grundsätzlich ist zu prüfen, ob das Verb mit der Partikel zusammen geschrieben werden kann oder nicht.

- Wir werden auch kommen $\Rightarrow$ Wir kommen auch/ADV
- Wir werden mitkommen $\Rightarrow$ Wir kommen mit/PTKVZ

Da die Getrennt- und Zusammenschreibung aber ihrerseits sehr uneinheitlich geregelt ist, liefert dieser Test meistens keine eindeutige Antwort. Ein weiterer Test ist die nicht-kompositionale Bedeutung. Wenn die Kombination von Verb+Partikel zu einer anderen als der kompositionalen Bedeutung führt, handelt es sich stets um PTKVZ:

- Wir bauen den Schrank zusammen/ADV (= gemeinsam)
- Wir bauen den Schrank zusammen/PTKVZ (= aus Einzelteilen)

Die Veränderung kann auch darin bestehen, daß sich der Valenzrahmen des Verbs ändert:

- Ich werde dich zur Königin machen.
- *Ich werde dich zur Königin weitermachen.

Also:

- Ich mache weiter/PTKVZ.

Insbesondere auf diesem Gebiet enthält das Lexikon viele Lücken. Wenn also ein zusammengesetztes Verb mit deutlicher Sinnänderung fehlt, ist es in `Verben.txt` einzutragen.

APPO oder PTKVZ?

Die Postpositionen 'herunter', 'herauf', 'herab', 'hinunter', 'hinauf', 'hinab' und 'entlang' sind immer APPO und niemals PTKVZ.

APZR oder PTKVZ?

Verschiedene Partikel können sowohl zur Präposition als auch zum Verb treten:

- Sie stürzte mit ausgebreiteten Armen auf ihn zu/APZR.
- Die Tür fiel zu/PTKVZ.
- Sie irrten vierzig Jahre lang herum/PTKVZ.
- Er segelte um den Kontinent herum/APZR.

Wenn die Partikel nicht mit dem Verb zusammengeschrieben werden kann, liegt APZR vor:

- Er kritisiert den Kompromiß von einer moralischen Position aus/APZR.
- *Der Kompromiß wird auskritisiert.

Wenn sowohl ein Verb als auch eine Präposition auftritt, zu der die Zirkumposition gehören könnte, und deutliche Bedeutungsunterschiede zu erkennen sind, entscheidet die wahrscheinlichere Satzbedeutung:

- Sie kam um die Säule herum/APZR (= trat vor).
- Er kam um die Wehrpflicht herum/PTKVZ (= mußte nicht dienen).
- Um 2000 kamen wir viel in Asien herum/PTKVZ (= machten viele Reisen).

Wenn beide Varianten dieselbe Bedeutung erzeugen, liegt APZR vor:

- Sie ging ohne Rührung an uns vorbei/APZR.

Wenn keine Präposition auftritt, liegt entweder PTKVZ oder ADV vor; vgl. den Abschnitt 'ADV oder PTKVZ?'.

FM, NN oder NE?

Bei Eigennamen, die eigentlich Ausdrücke einer Fremdsprache sind, ist es schwer zu entscheiden, ob es sich bei den Bestandteilen um Namen handelt. Wo eine ganze Phrase einer fremden Sprache als ein zusammengesetzter Name behandelt wird, sollte idealerweise die gesame Phrase als *ein* Wort behandelt werden, das dann die Katgorie NE hat:

- Die Gruppe 'Frankie goes to Hollywood'/NE entstand 1979.

Wenn der Name aber in Form von mehreren Worten im Satz auftritt, weil der Tokenizer ihn nicht kannte, soll jedes Wort so klassifiziert werden, wie es auch alleinstehend klassifiziert würde:

- Die Gruppe 'Frankie/NE goes/FM to/FM Hollywood/NE' entstand 1979.

Ein Wort wie 'to' wird also nicht als Eigenname angesehen, nur weil es in einem Namen vorkommt, aber 'Hollywood' wäre auch sonst ein Eigenname. Bisweilen können auch noch andere Kategorien in einem zusammengesetzten Namen vorkommen:

- Das brandneue MacOS/NE X/CARD Server/FM (= das Betriebssystem) kostet 499$.
- Der brandneue MacOS/NE X/NE Server/NN (= das X11-Fenstersystem) ist kostenlos.

'Mac OS X Server' ist eine Variante des zehnten Betriebssystems MacOS, also handelt es sich bei 'X' um eine Zahl und bei 'Server' um ein Fremdwort (sonst müßte es 'der Server' heißen). Im zweiten Satz ist dageegen von einer Implementation des Servers mit dem Namen 'X' für das Betriebssystem MacOS die Rede, also ist 'X' ein Eigenname (im Besitz des M.I.T.) und 'Server' ein normales Nomen.

PRELS oder PWS?

Das Wort 'was' kann sowohl Frage- als auch Relativpronomen sein. In der direkten Frage ist es Fragepronomen und im Relativsatz mit Satzreferenz Relativpronomen:

- Was/PWS soll das bedeuten? Es taget ja schon!
- Es tagt schon, was/PRELS ich sehr merkwürdig finde.

Bei Relativsätzen, die sich auf NP beziehen, ist zu prüfen, ob sie aus der betreffenden Gesamtfrage hervorgegangen sind oder ob das 'was' semantisch leer ist:

- Ich kann das, was/PRELS geschehen ist, nicht ungeschehen machen.
  (impliziert *nicht* die Frage 'Was ist geschehen?')
- Sie fragte ihn, was/PWS geschehen sei.
  (= Sie fragte: "Was ist geschehen?")

VAFIN oder VVFIN?

Die Verben 'sein', 'haben' und 'werden' werden immer als Auxiliarverben eingeordnet, egal ob sie mit oder ohne anderes Verb stehen.

- Ich bin/VAFIN, der ich bin/VAFIN.
- Wir sind/VAFIN lange umhergezogen.
- Wir sind/VAFIN müde.

VMINF oder VMPP?

Die Formen 'wollen', 'können' etc. werden auch dann als Infinitiv eingeordnet, wenn sie syntaktisch die Rolle eines Partizips einnehmen.

- Er hat nicht kommen können/VMINF.
  (= Er hat nicht zu kommen vermocht/VVPP.)

VVFIN oder VVIMP?

Plural- oder Infinitivformen von Verben werden stets nach ihrer morphologischen Struktur eingeordnet, *nicht* als Imperative, auch wenn sie eindeutig imperative Bedeutung haben.

- Gehe/VVIMP über Los!
- Gehen/VVFIN Sie nicht über Los!
- Stehenbleiben/VVINF!

1.2.4 Welches Label?

ADV oder AVZ?

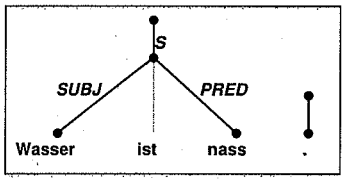
AVZ wird für alle Worte der Kategorie PTKVZ verwendet und ADV für Worte der Kategorie ADV; diese Entscheidung ist also zurückzuführen auf die Frage 'ADV oder PTKVZ?' (siehe oben).

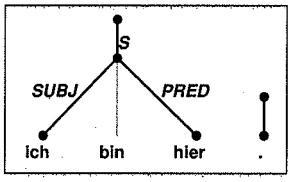
ADV oder PP?

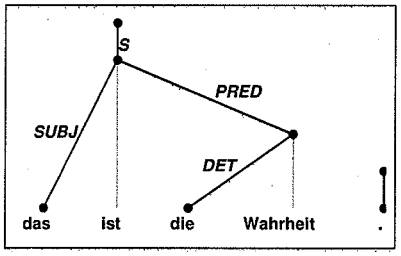
Präpositionaladverben wie 'dazu', 'hiermit' etc. sind Verschmelzungen von Präposition und Adverb. Logischerweise könnten sie also sowohl ADV als auch PP sein. Der Einheitlichkeit halber wird immer PP verwendet.

ADV oder PRED?

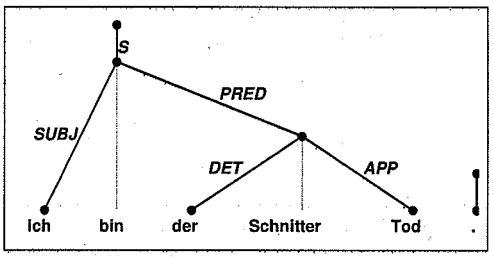
Als Prädikativum (PRED) wird nur ein Komplement bezeichnet, das in etwa eine Gleichsetzung ausdrückt:

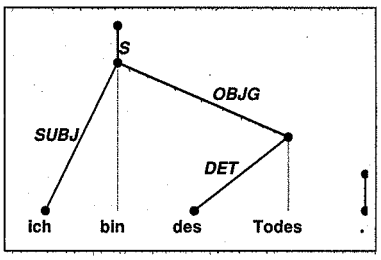




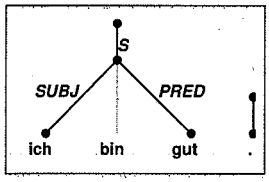


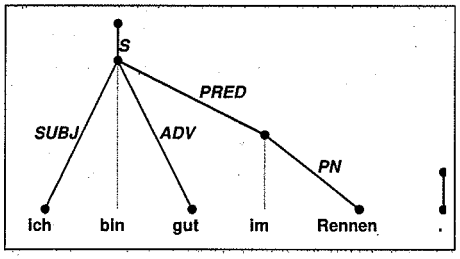
Es gibt einige Tests, die das Prädikativum von Objekten, adverbialen Bestimmungen etc. unterscheiden. Zum Beispiel muß es immer im Nominativ stehen:

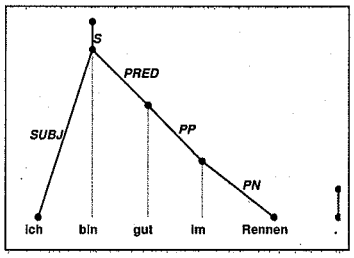




Es alterniert mit den anderen Komplementen von 'sein':

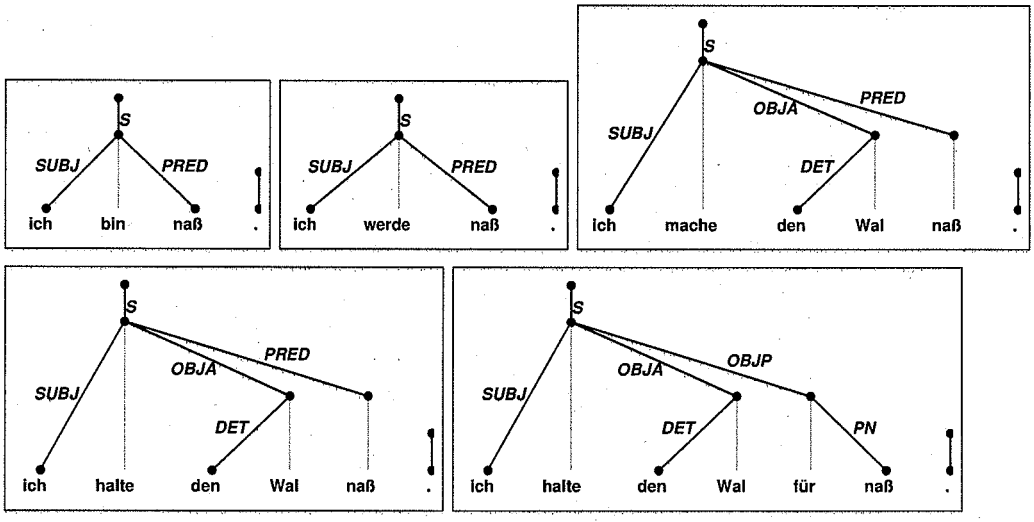






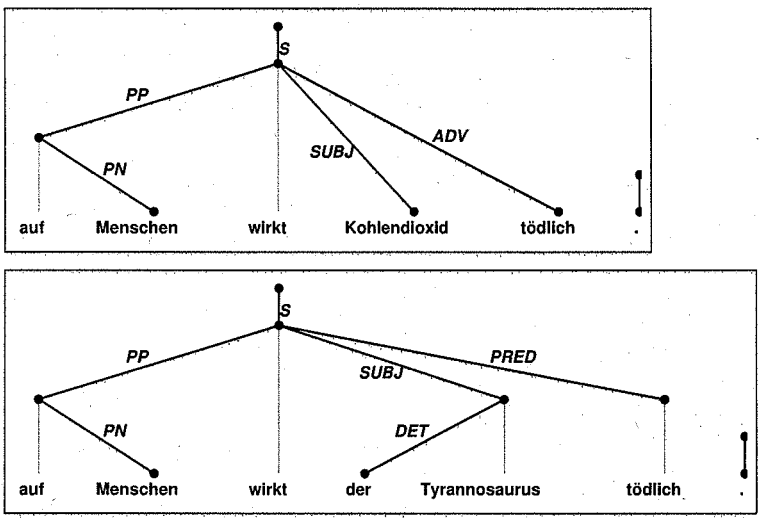
Die zweite Konstruktion drückt sinngemäß aus, daß man in Führung liege, die dritte, daß man gut rennen kann.

Dennoch ist die Grenze fließend:



Wir erlauben das Label PRED an verschiedenen anderen Verben außer 'sein'. Richtlinie ist dabei, daß das Verb die elementare Aussage 'A ist B' ausdrückt oder sie zumindest impliziert. Daher gilt 'den Wal naß machen' als Prädikativum (denn es impliziert 'der Wal ist naß'), während 'schön singen' nichts dergleichen impliziert (der talentierte Sänger kann häßlich sein wie die Nacht).

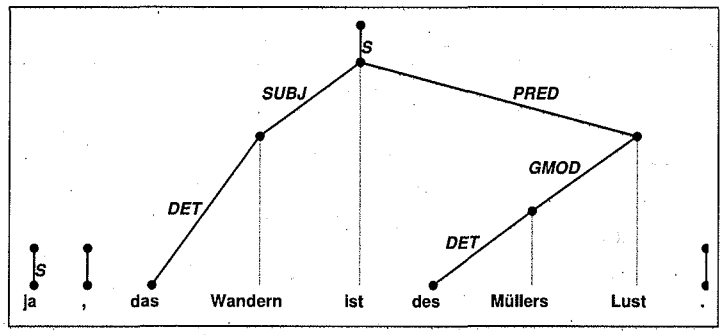
Die Gleichsetzung kann auch nur im Geiste des Sprechers stattfinden:

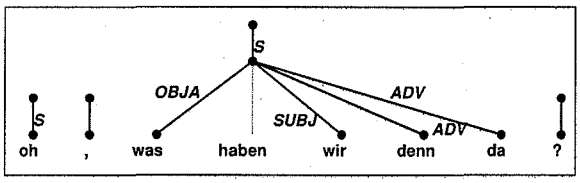


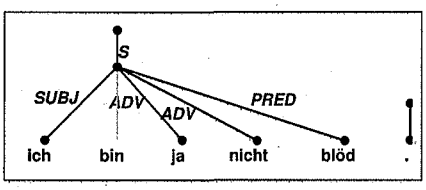
Das farb- und geruchlose Kohlenmonoxid ist beim Einatmen tödlich, aber es *wirkt* (= erscheint) dem Menschen nicht tödlich. Daher ist ADV zu annotieren. Dagegen ist das meterlange Gebiß des Tyrannosaurus deutlich zum Töten bestimmt, bewirkt also im Betrachter den Eindruck "Der Saurier war tödlich". Deshalb ist PRED zu annotieren.

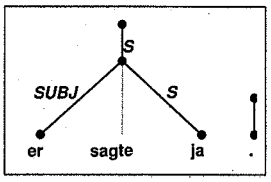
ADV oder S?

Wirkliche Interjektionen (PTKANT, ITJ) sind stets S. Sie sind immer Satzwurzeln, außer es gibt ein übergeordnetes Aussageverb. Nur wenn die Wortstellung deutlich macht, daß es sich um ein adverbial gebrauchtes Wort der Kategorie ADV handelt, ist ADV zu verwenden.



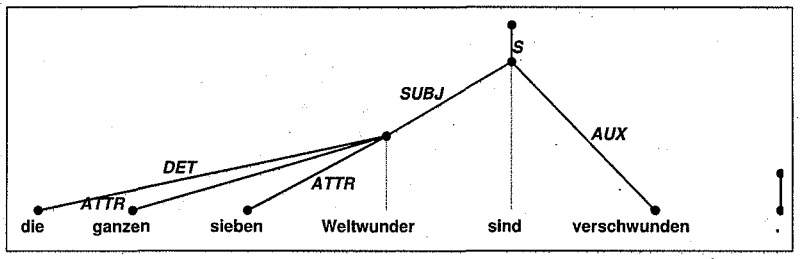




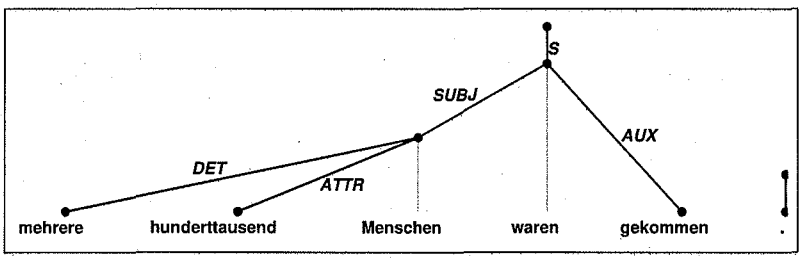


APP oder DET?

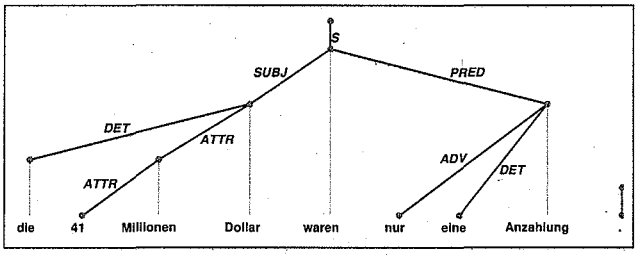
Wenn Zahlausdrücke aus mehreren Worten bestehen, werden alle adjektivischen Bestandteile als ATTR oder DET dem Nomen untergeordnet.



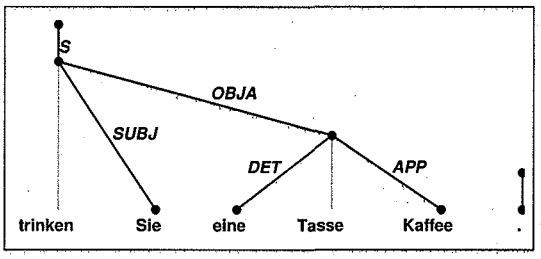
Das gilt auch dann, wenn inhaltlich das eine Zahlwort das andere modifizieren könnte:



Wenn Zahlworte aber syntaktisch Nomen sind (wie 'Millionen' etc.), sind sie ATTR zum Hauptnomen. Vorangehende Bestandteile werden dem Zahlnomen untergeordnet.

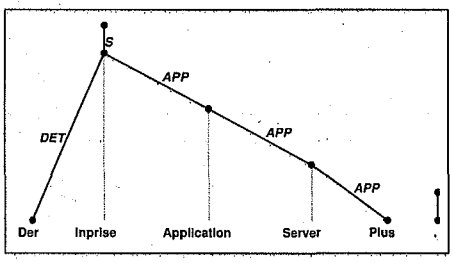


Dagegen gelten Kombinationen von Maßnomen+Stoffnomen als Apposition:

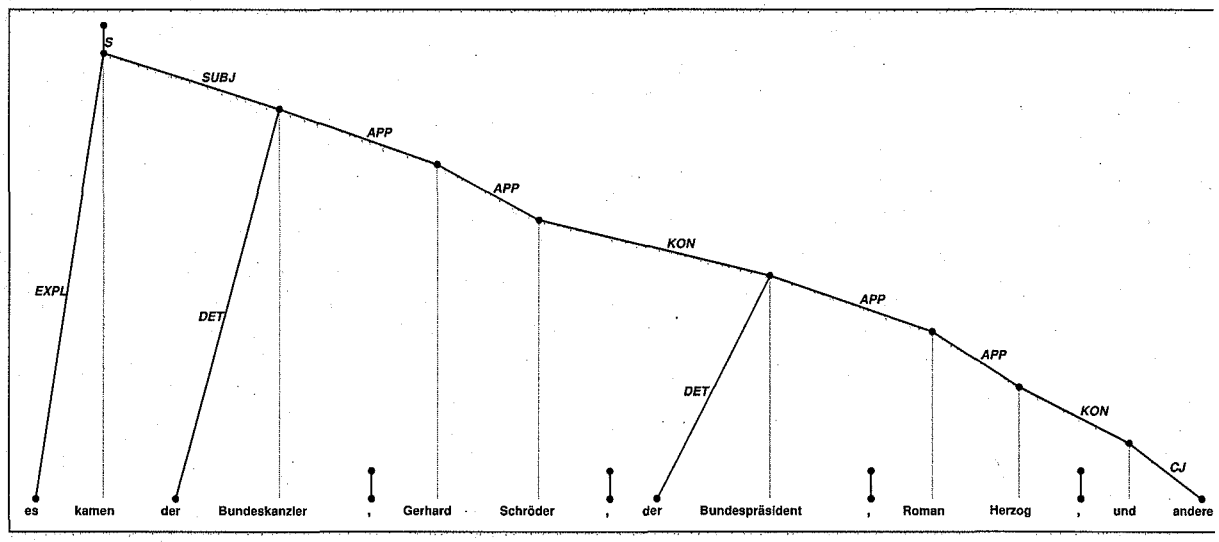


APP oder KON?

In komplexen Nominalphrasen kommt gewöhnlich APP zum Einsatz, egal wieviele Bestandteile sie haben.



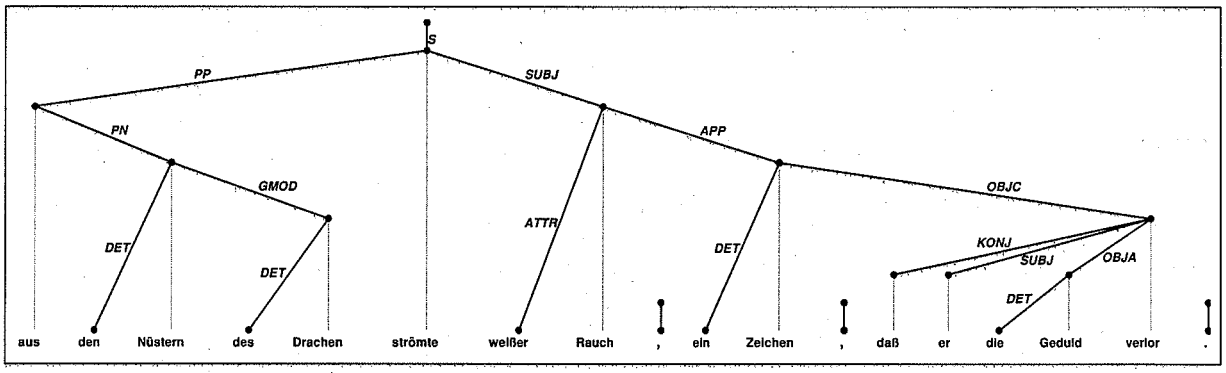
KON wird nur verwendet, wenn eindeutig verschiedene Referenten bezeichnet werden:



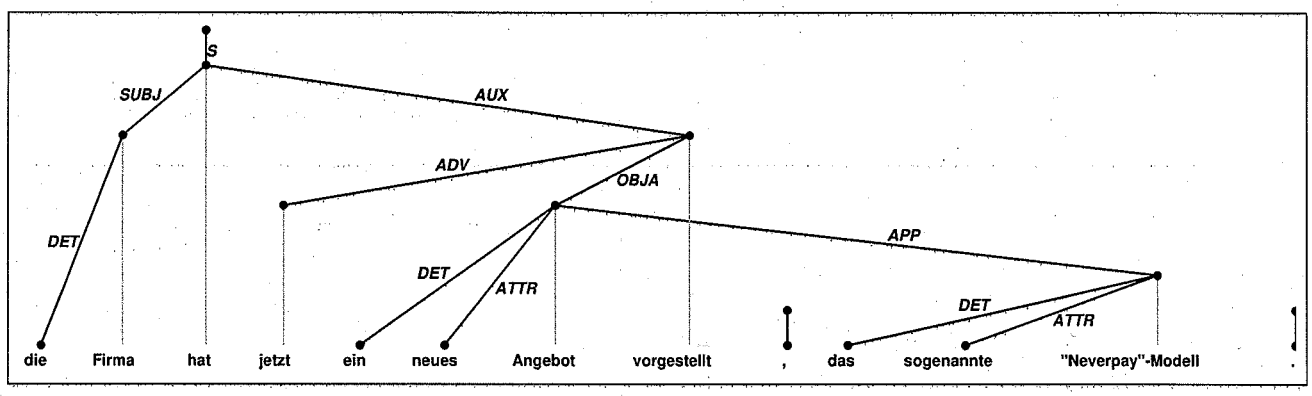
APP oder S?

Nachgeschobene NP sind manchmal als Erläuterungen einer früheren NP aufzufassen, manchmal aber auch als Erläuterungen eines ganzen Satzes.

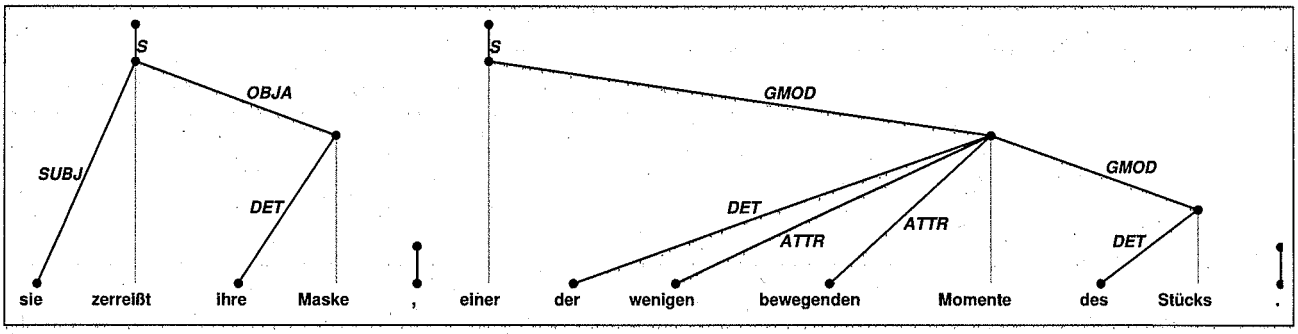
Wenn das Bezugswort eindeutig ein vorhergehendes Nomen ist, dann modifiziert die nachgeschobene NP dieses Nomen als APP:



Das gilt auch dann, wenn die Unterordnung nichtprojektiv ist:

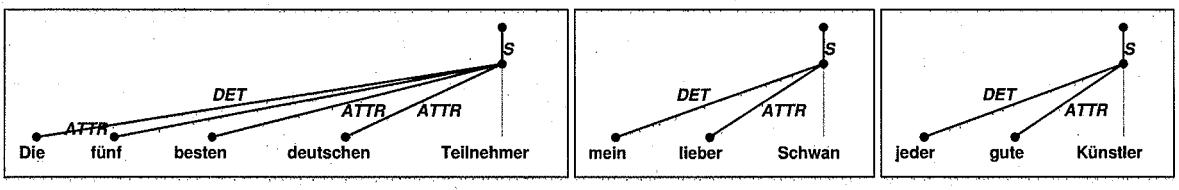


Wenn hingegen ein Nomen die Aussage eines ganzen Satzes aufgreift oder wiederholt, ist es als Fragment eines zweiten Satzes aufzufassen:

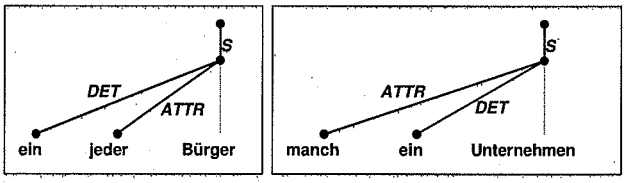


ATTR oder DET?

DET wird für Artikel und attributive Pronomen gebraucht. Alle anderen Bestandteile von NP, die nicht Nomen im weiteren Sinne sind (NN, NE, FM, TRUNC) sind als ATTR zu bezeichnen.



Wenn sowohl Artikel als auch attributives Demonstrativpronomen stehen, ist der Artikel DET und das Pronomen ATTR.

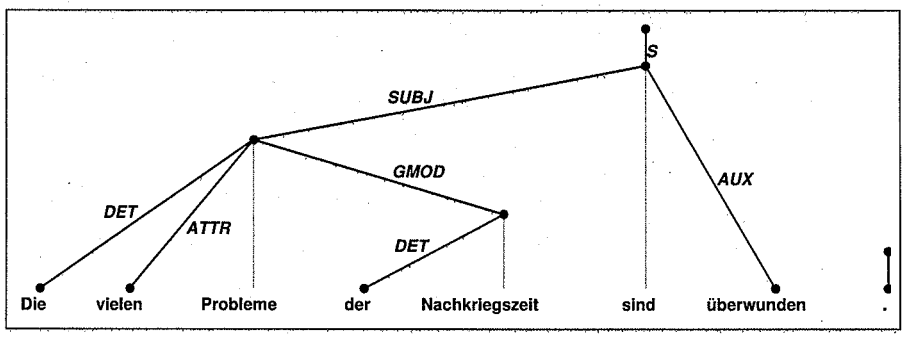


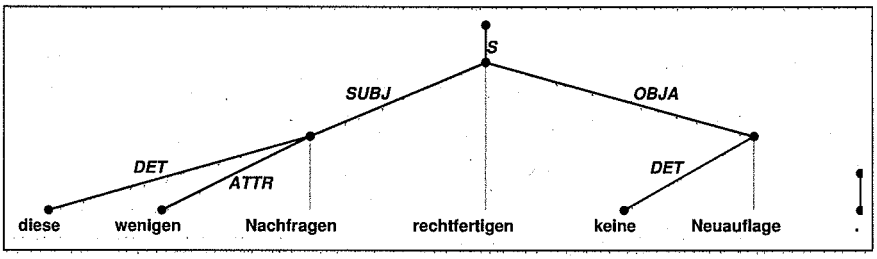
Treten zwei attributive Pronomen auf, so sind sie als Beiordnung anzusehen:

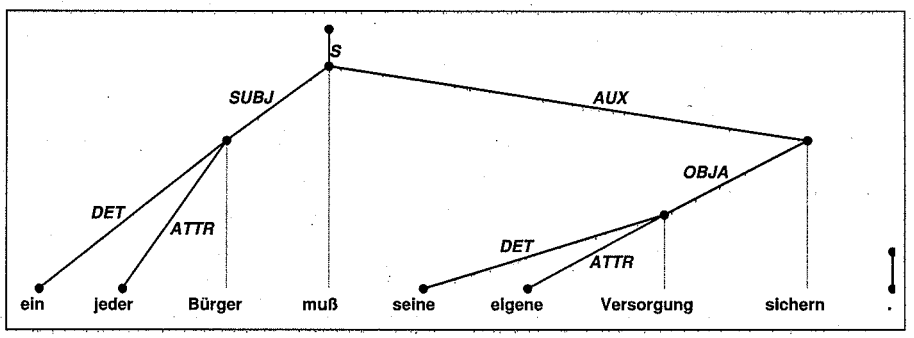
- dieses/DET unser/KON Land

Bei Pronomen, die sowohl mit als auch ohne Artikel stehen können, hängt die Bezeichnung sowohl vom einzelnen Wort ab als auch davon, ob ein weiterer Determiner auftritt.

Steht neben dem Pronomen ein weiteres Pronomen oder Artikel, das Determiner ist, so ist immer ATTR zu wählen:



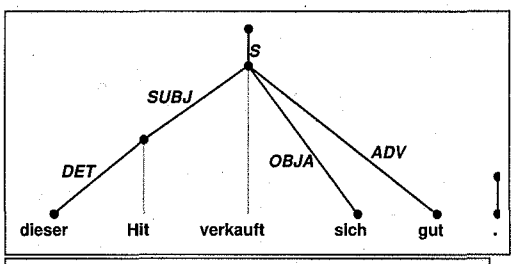


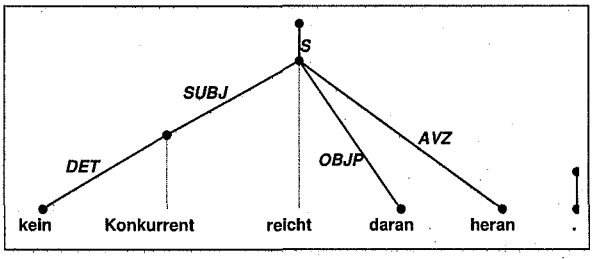


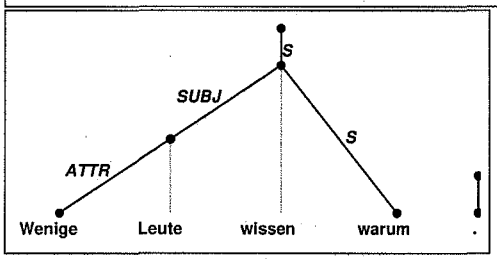
Wenn ein Pronomen regelmäßig neben stark oder gemischt gebeugten Adjektiven auftritt, ist es DET; wenn ein Pronomen neben schwach gebeugten Adjektiven auftritt, ist es ATTR:

- Dieser/DET neue/weak Hit verkauft sich gut.
- Diese/DET neuen/weak Hits verkaufen sich gut.
- Die Single ist kein/DET großer/mixed Erfolg.
- Die Singles sind keine/DET großen/mixed Erfolge.
- Wenig/ATTR neuer/strong Inhalt ist darauf.
- Wenig/ATTR neue/strong Inhalte sind darauf.

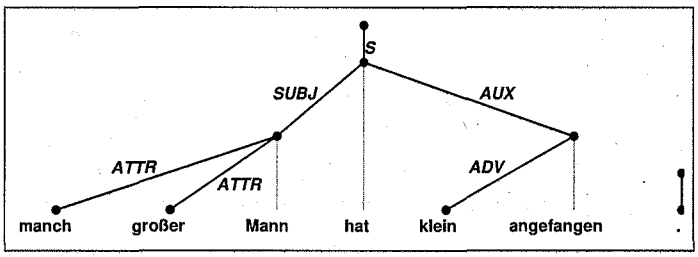
Also sind 'dieser' und 'keiner' stets DET, und 'wenig' ist immer ATTR, auch dann wenn sie allein stehen, die Flektionsstufe also nicht erkennbar ist:

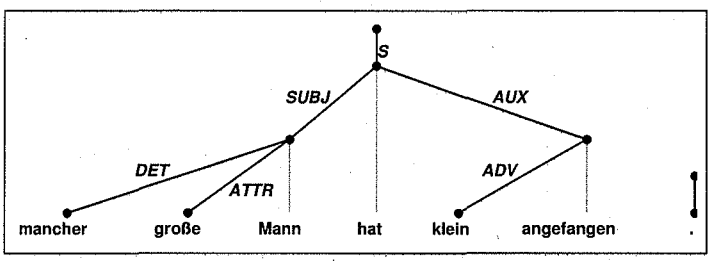






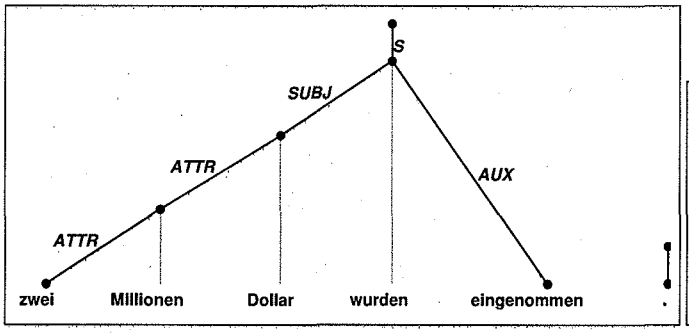
Hierbei verhalten sich die einzelnen Pronomen ganz unterschiedlich. So sind etwa die Pronomen 'manch' und 'welch' immer ATTR, während 'mancher' und 'welcher' immer DET sind:

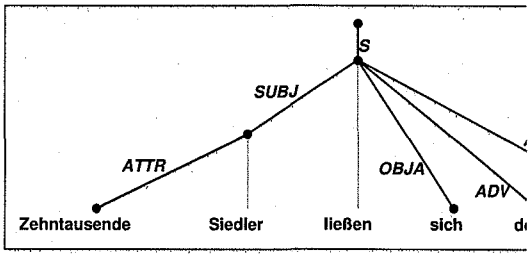




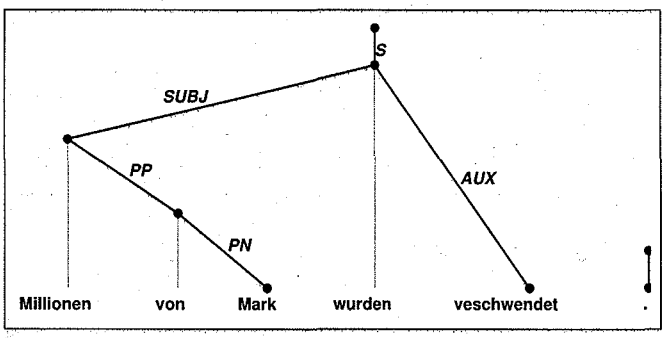
ATTR oder SUBJ?

Zahlworte, deren Kategorie NN ist, werden als ATTR angesehen.

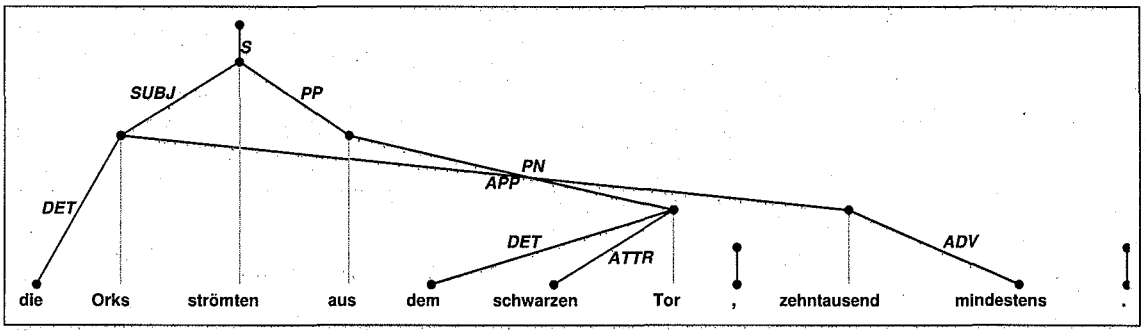




Wenn Zahlwort und Nomen durch Präposition getrennt sind, ist jedoch das Nomen PN.



Ist das Zahlwort nachgestellt, so ist es stattdessen APP.

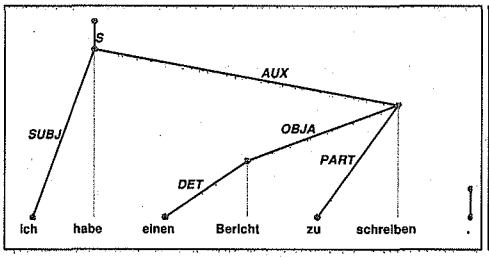


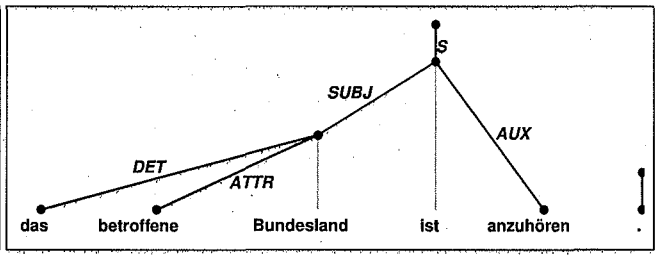
AUX oder OBJI?

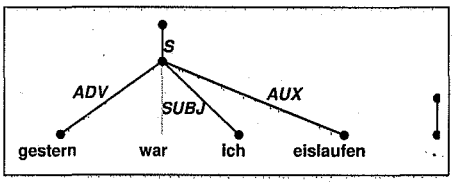
Untergeordnete Infinitive können sowohl AUX als auch OBJI sein.

- Pseudo-Auxiliarphrasen mit 'sein' und 'haben'

  Die Hilfsverben 'sein' und 'haben' bilden normalerweise Auxiliarphrasen mit dem Partizip Passiv. Sie können aber auch Infinitive nehmen wie das Wort 'werden' und mit ihnen eine Art Auxiliarphrase bilden. Auch in diesem Fall werden Infinitiv und Hilfsverb durch AUX verbunden.



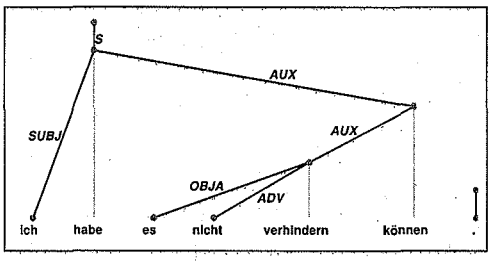


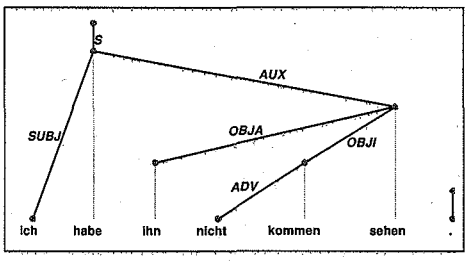


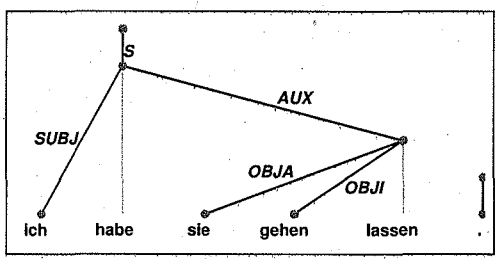
Diese Konstruktionen sind als Auxiliarphrasen anzusehen, weil sie im wesentlichen nur die Satzaussage mit Aspekt-, Zeit- oder Modalinformation anreichern: 'Ich habe einen Bericht zu schreiben' impliziert 'Ich schreibe einen Bericht' oder zumindest 'Ich muß einen Bericht schreiben'; 'Das Bundesland ist anzuhören' bedeutet 'Das Bundesland wird angehört' oder zumindest 'Das Bundesland soll angehört werden'; und 'ich war eislaufen' impliziert 'ich lief eis'.

- 'haben' mit Pseudo-Infinitiv

  Mit einigen Verben bildet das Verb 'haben' das Perfekt sogar regelmäßig mit dem Infinitiv statt mit dem Partizip Passiv. Es handelt sich um die Modalverben 'müssen', 'können', etc, außerdem Wahrnehmungsverben wie 'sehen' oder 'hören' und auch das Verb 'lassen'. In allen diesen Fällen ist AUX zu annotieren, obwohl das untergeordnete Verb die Kategorie VMINF oder VVINF hat.

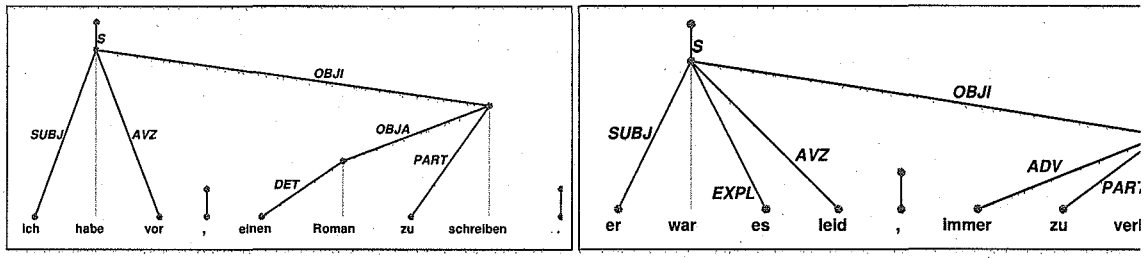




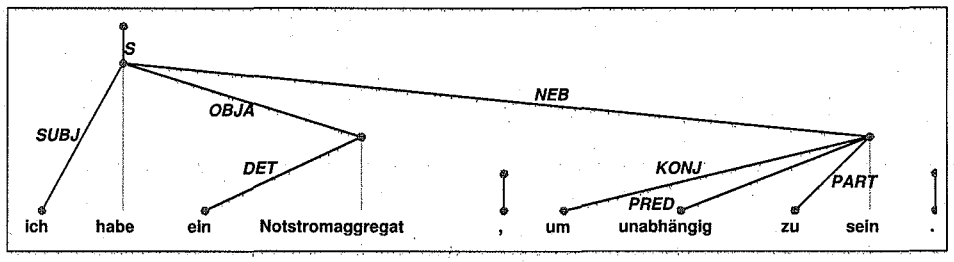


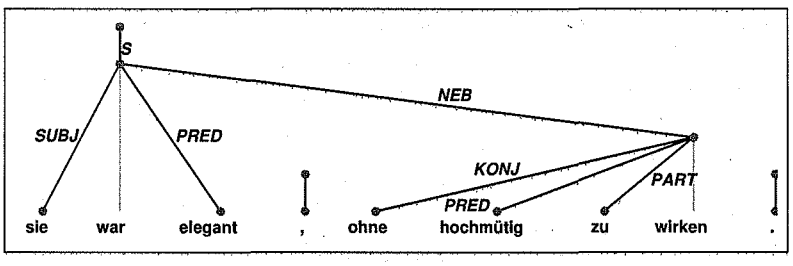
- Auxiliarverben als Vollverben

  'Sein' und 'haben' können aber auch als Vollverben auftreten und dann eine zweite, zumindest teilweise unabhängige Satzaussage unterordnen. In diesem Fall ist nicht AUX zu wählen. Wenn zum Beispiel ein trennbares Verbpräfix dazutritt, ist OBJI zu wählen.



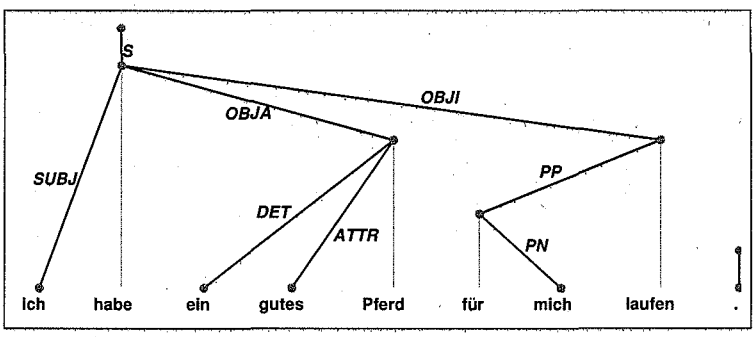
Handelt es sich um einen erweiterten Infinitiv mit 'um zu' oder ähnlichen Konstruktionen, ist NEB zu wählen.

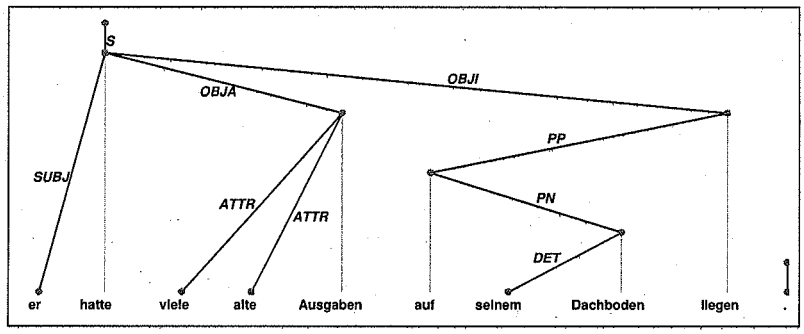




- 'Ein Experiment laufen haben'

Das Verb 'haben' kann auch eine Konstruktion mit Nominal- und Verbalobjekt eingehen, die *nicht* den einfachen Satz implizieren:

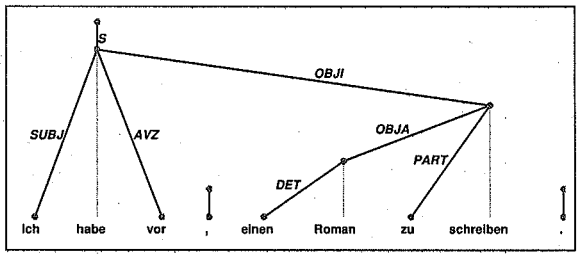


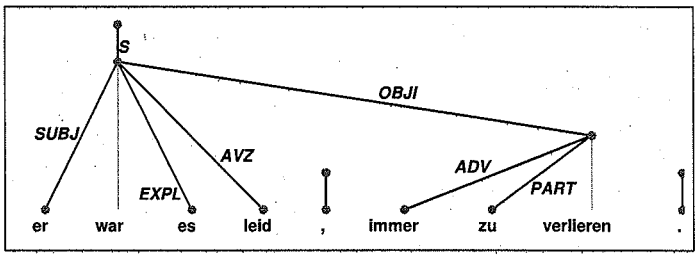


'Ich habe ein Pferd laufen' impliziert nicht 'Ich laufe ein Pferd', sondern sowohl 'Ich habe ein Pferd' als auch 'das Pferd läuft'. Daher ist 'laufen' eigener Satz(infinitiv) anzusehen, der mit OBJI bezeichnet wird.

- am Vollverb

  Wenn Infinitive am Vollverb stehen, sind sie niemals AUX, sondern stets OBJI.

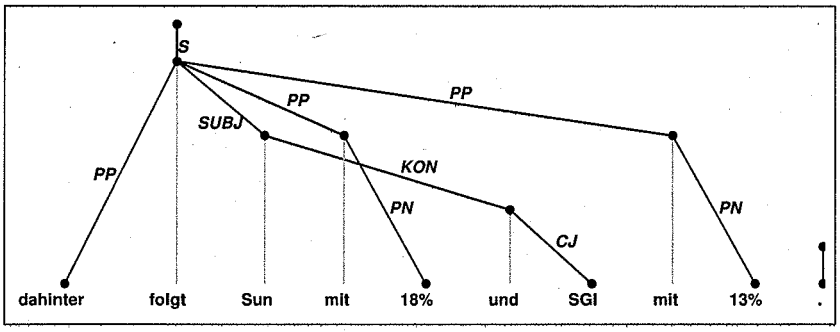




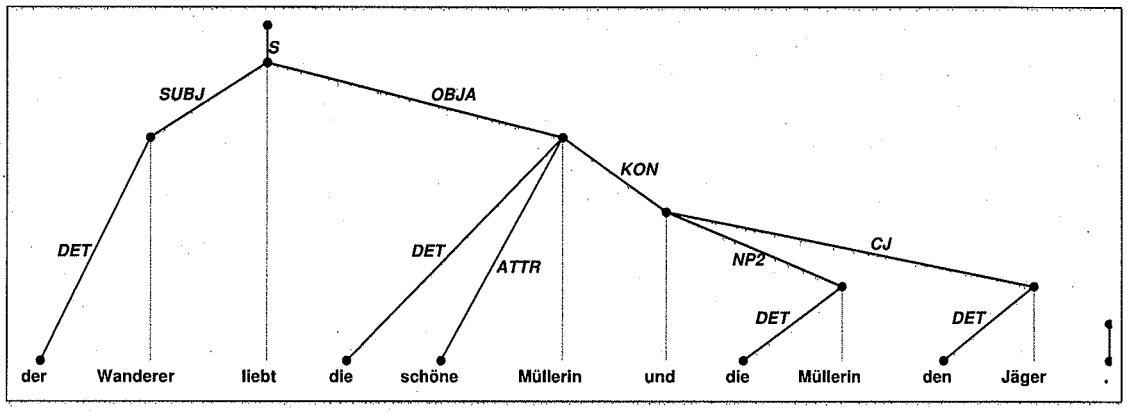
Das gilt auch für die Vollverben, die den Hilfsverben in ihrer Funktion sehr ähnlich sind ('lassen', 'scheinen', 'brauchen').

CJ oder NP2?

Das Label NP2 für ausdrücklich elliptische Koordinationen mit zweitem Subjekt sollte nur dann verwendet werden, wenn andere Konstruktionen nicht möglich sind.



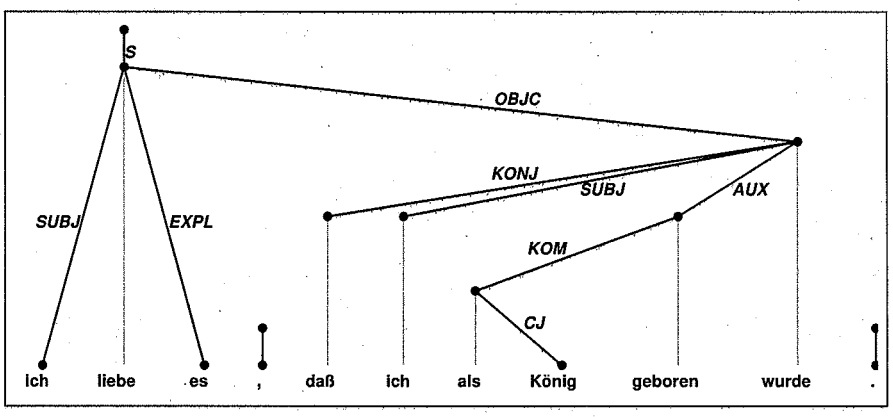
('SGI mit 13%' ist eine mögliche und sinnvolle NP.)

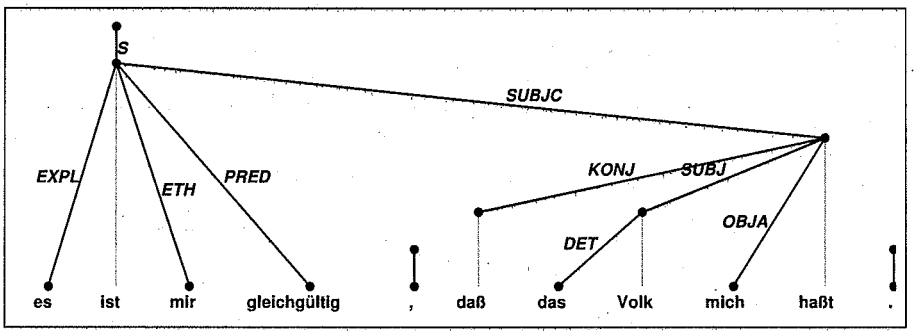


('Die Müllerin den Jäger' ist keine NP.)

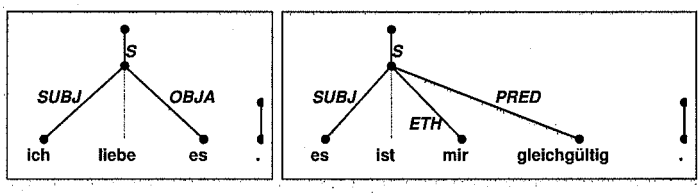
EXPL oder OBJA?

Steht das 'es' vor einem Objektsatz oder Subjektsatz, der zum vorigen Verb gehört, so ist es EXPL.

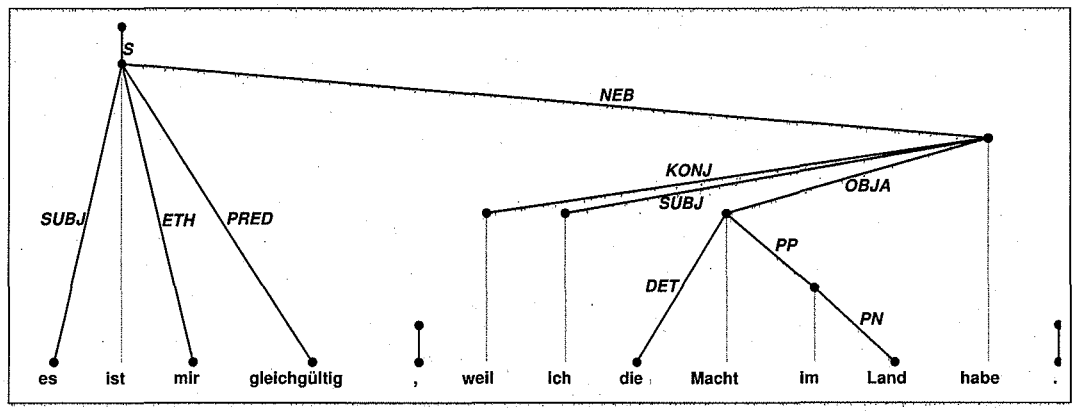




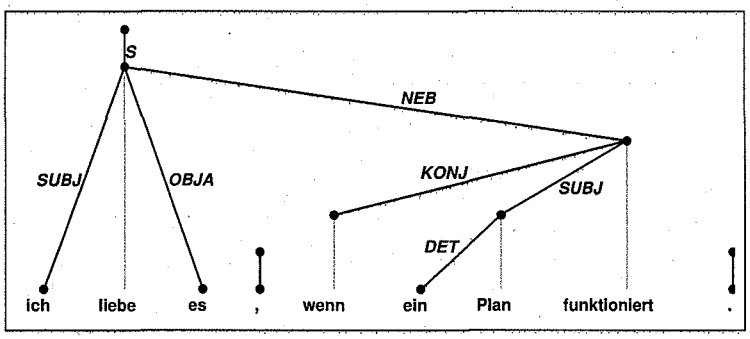
Wenn kein solcher Satz auftritt, ist das 'es' normales Subjekt oder Objekt.

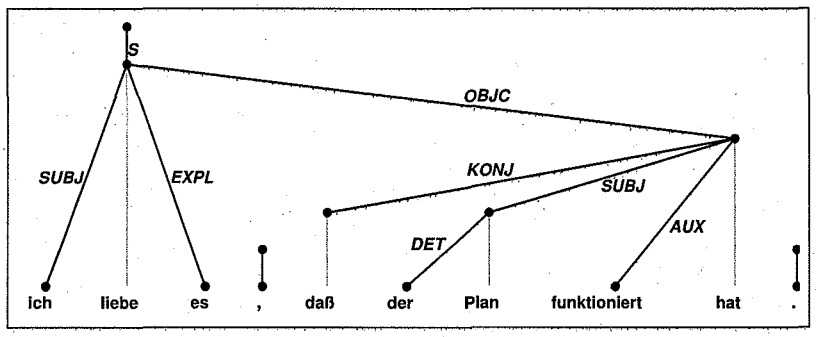


Wenn nur ein normaler Nebensatz folgt, so ist 'es' SUBJ oder OBJA.



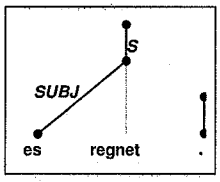
Auch wenn der Nebensatz inhaltlich die Funktion des Objektsatzes übernimmt, ist dennoch OBJA zu wählen.





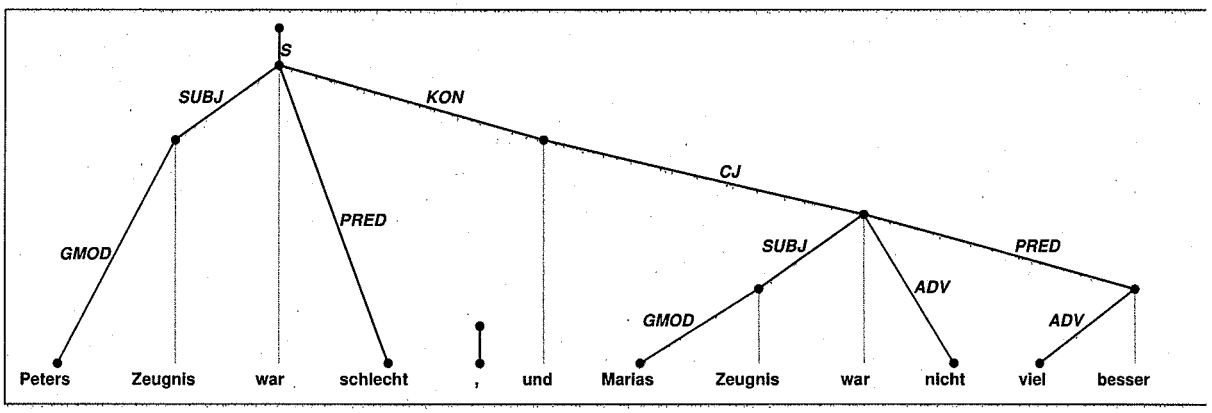
EXPL oder SUBJ?

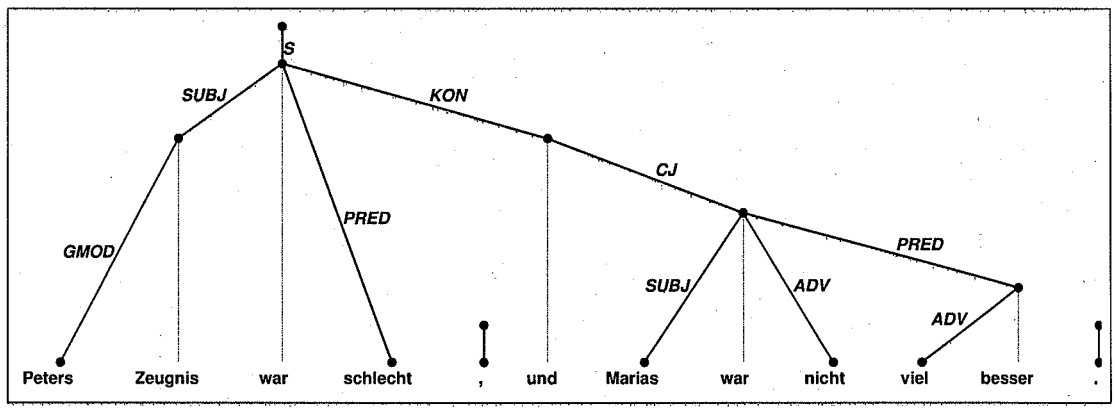
'es' bei unpersönlichen Verben ist stets SUBJ, nicht EXPL.



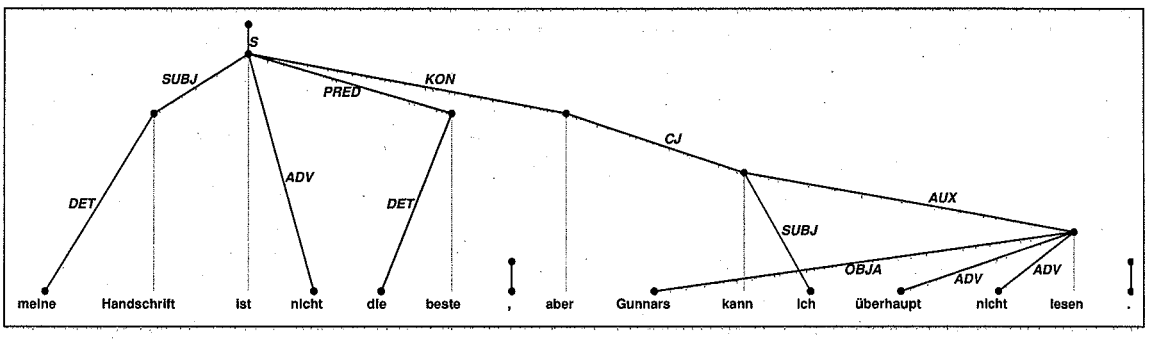
GMOD oder SUBJ?

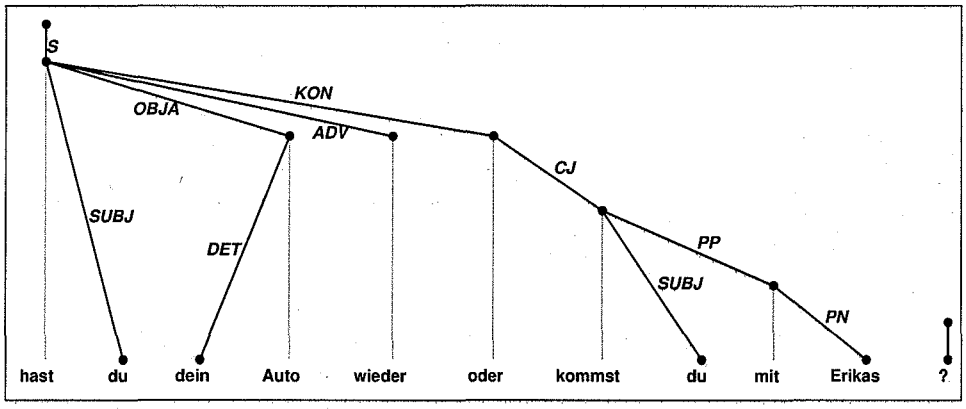
Nominalphrasen mit einem Eigennamen im Genitiv werden manchmal auf diesen Namen verkürzt. In diesem Fall übernimmt der Eigenname die Funktion des fortgefallenen Wortes.





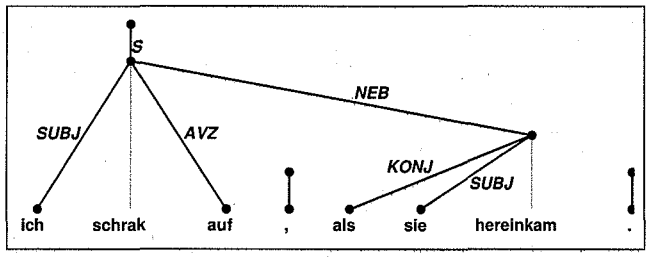
Das kann nicht nur die Funktion SUBJ sein, sondern z.B. auch OBJA oder PN.

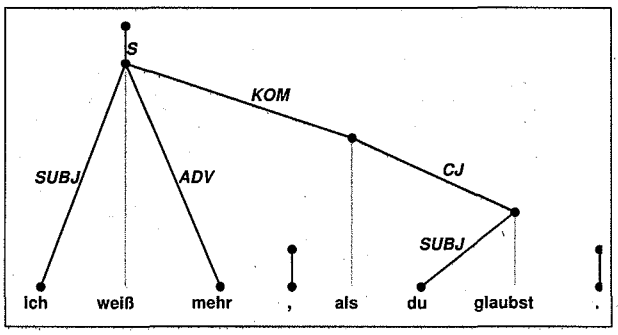




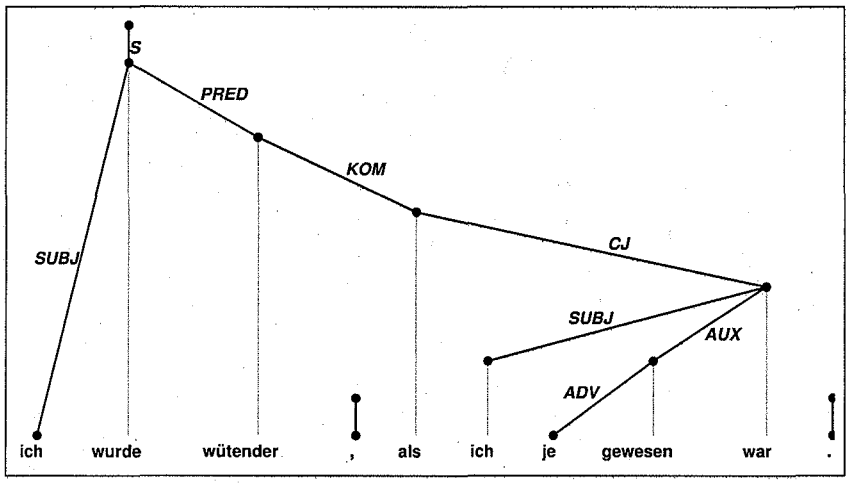
KOM oder KONJ?

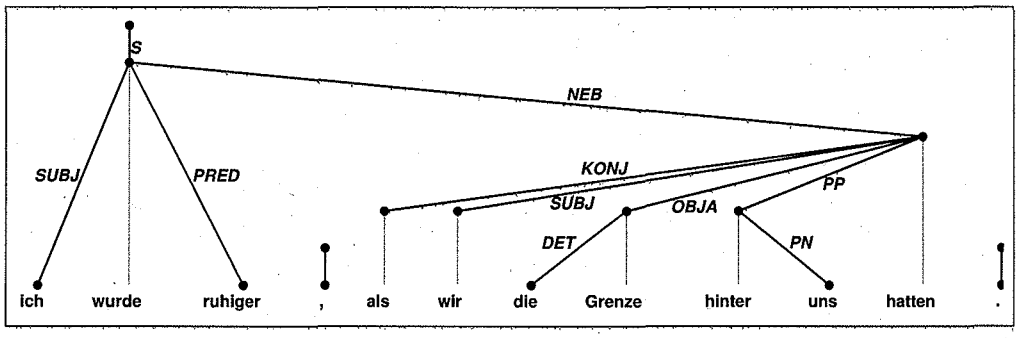
Das Wort 'als' kann Worte vieler Kategorien beiordnen, fast wie eine Konjunktion. Wenn es einen ganzen Nebensatz einleitet, kann es sowohl über- als auch untergeordnet werden. Ist mit 'als' ein Vergleich ausgedrückt, so wird es als KOM dem Hauptsatz untergeordnet. Ist Gleichzeitigkeit ausgedrückt, so wird es als KONJ dem Nebensatzverb untergeordnet.



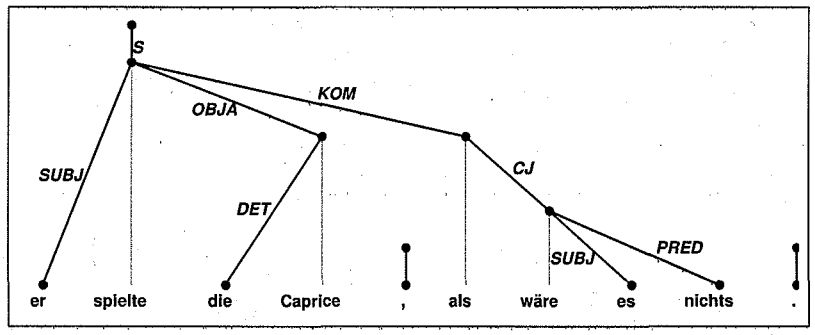


Die Vergleichsbedeutung tritt nur mit Komparativen auf; umgekehrt kann aber ein Hauptsatz mit Komparativ auch einen als-Nebensatz tragen.



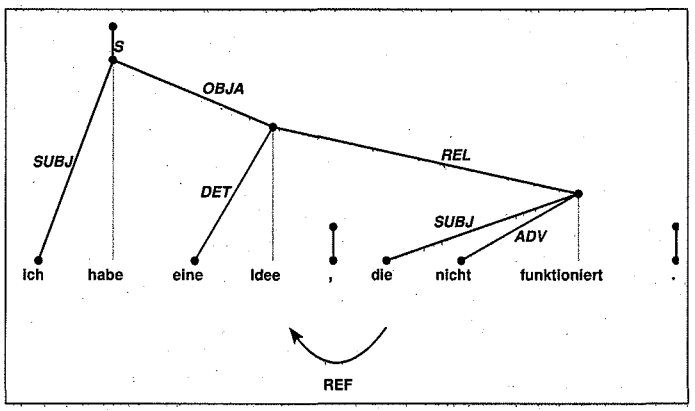


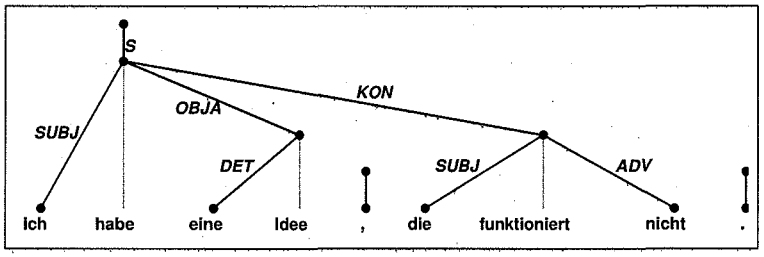
Wenn der Nebensatz in Verberststellung steht, ist 'als' immer KOM.



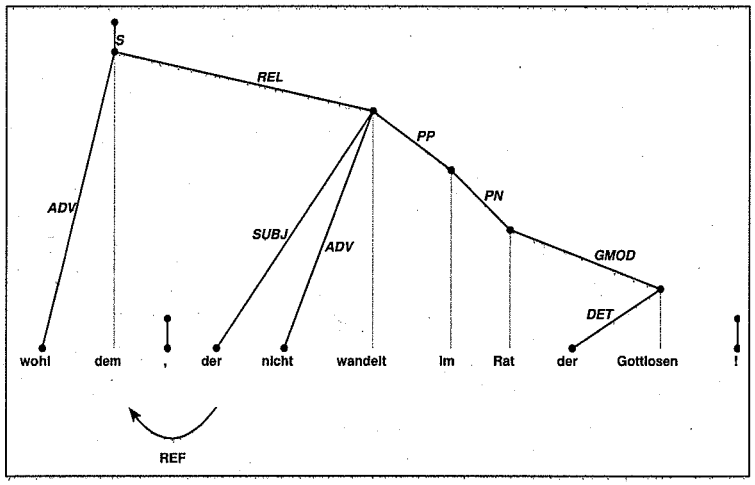
KON oder REL?

Beigeordnete kurze Hauptsätze mit Demonstrativpronomen als Subjekt ähneln oft Relativsätzen. Meistens kann die Wortreihenfolge enscheiden, ob Verbletztstellung vorliegt oder nicht:

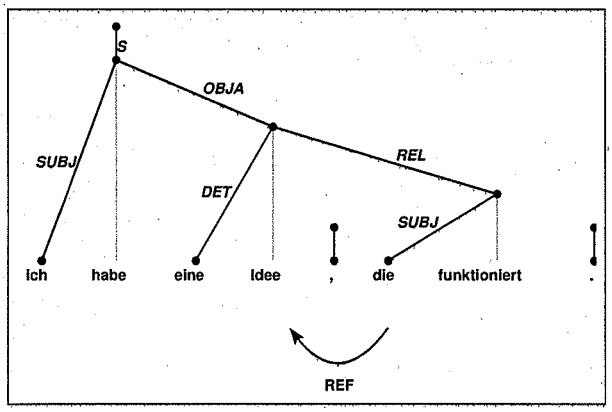




Wenn allerdings ein Text in altem Deutsch vorliegt, das die Verbletztstellung ganz allgemein nicht befolgt, ist auch entgegen der Wortstellung ein Relativsatz anzunehmen:

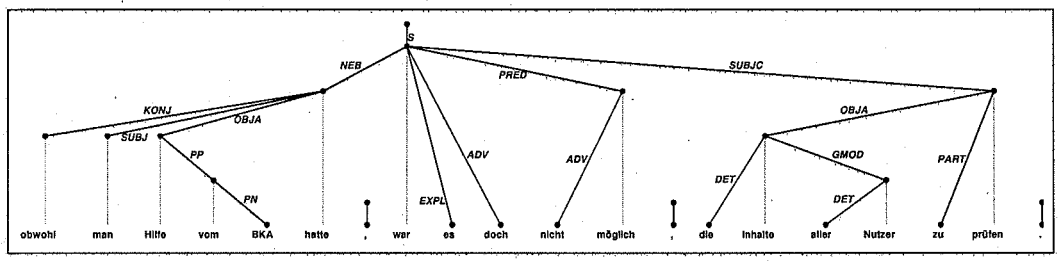


Wenn keine andere Modifikation auftritt als das Subjekt, läßt sich dieser Test nicht ausführen. Gewöhnlich ist ein solcher kurzer Satz als Relativsatz anzusehen:

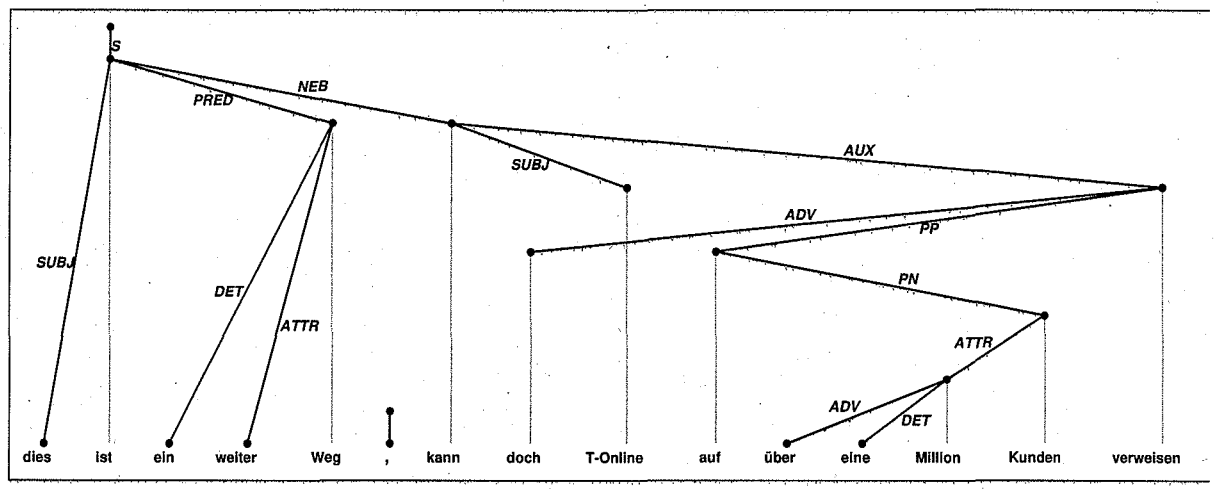


NEB oder KON?

Sätze, die mit 'doch' gebildet sind, können entweder normale Hauptsätze oder konjunktionslose Nebensätze sein. Wenn ein solcher Satz ein normales Vorfeld hat, ist er als S zu bezeichnen:



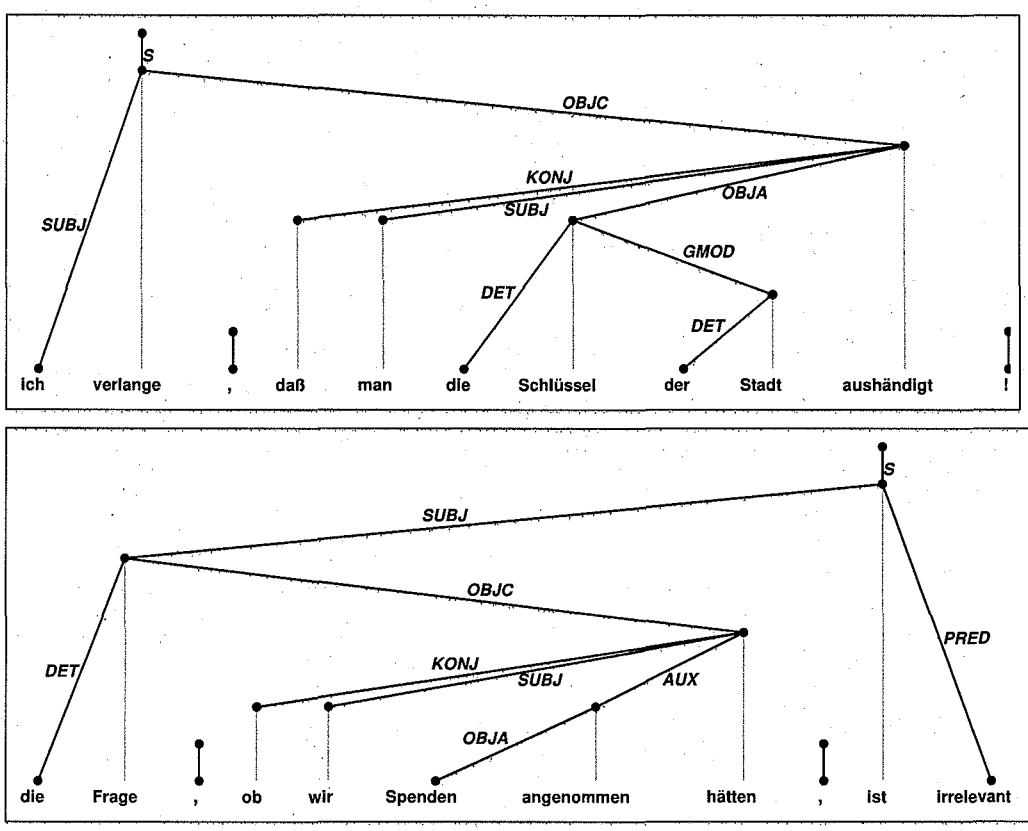
Wenn kein Vorfeld vorliegt und der Satz mit 'doch' die Funktion eines Kausalsatzes hat, ist NEB zu verwenden:



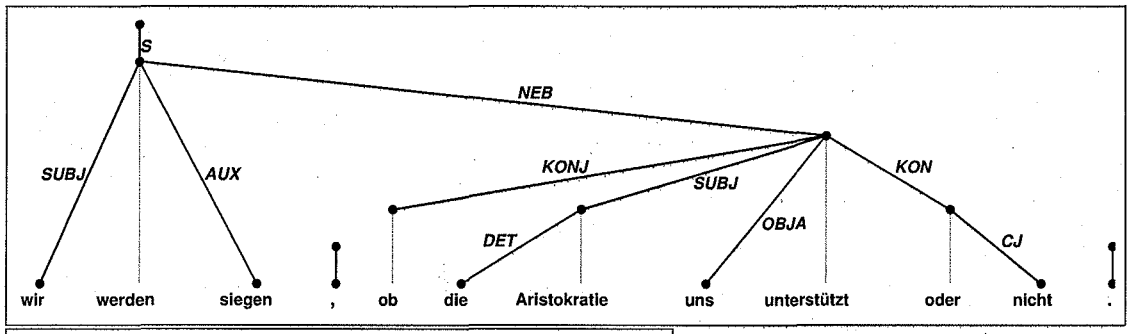
(= weil T-Online ... verweisen kann.)

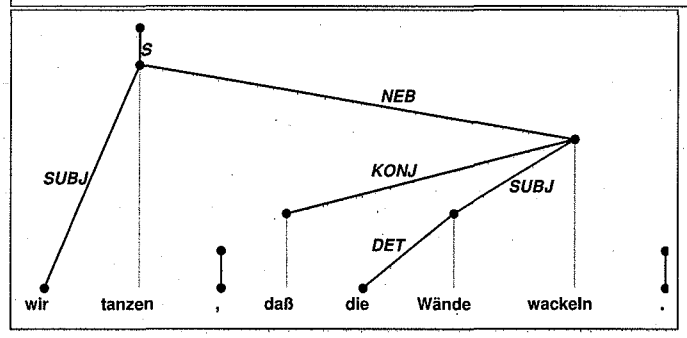
NEB oder OBJC?

Das Label OBJC wird nur vergeben, wenn der Regent diese Funktion verlangt. Das sind nur einige Verben, deverbale Nomen und Adjektive.

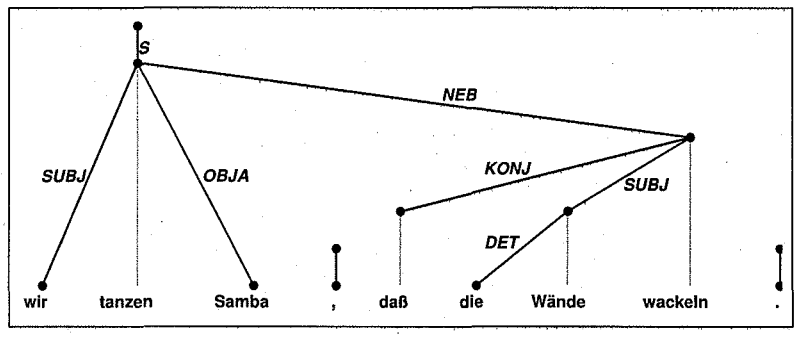


Wenn der untergeordnete Satz als Paraphrase von 'so daß' oder 'egal ob' anzusehen ist, ist er stattdessen NEB:





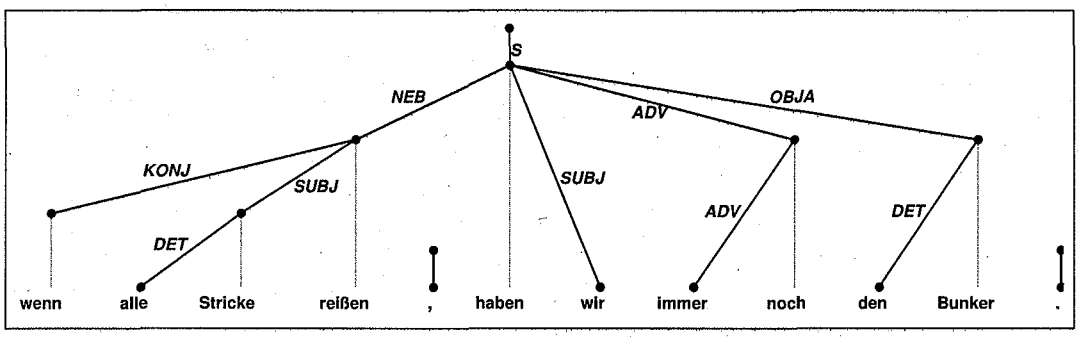
Deutliches Indiz für NEB ist, daß zu dem Satz noch das Objekt treten kann:

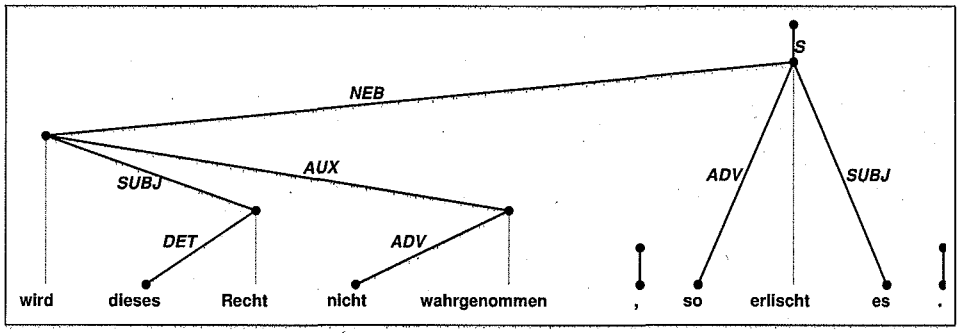


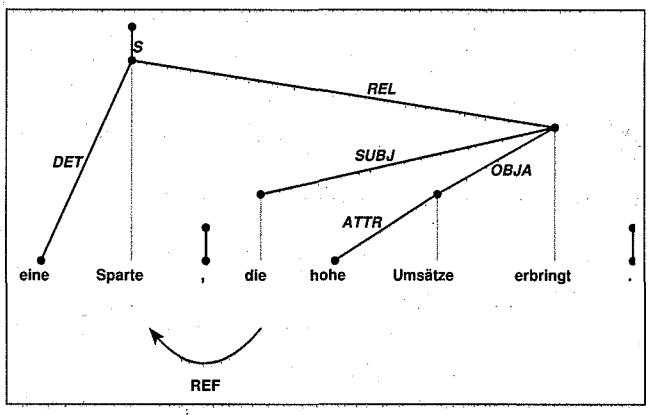
- *Wir verlangen eine Abrüstung/OBJA, daß abgerüstet wird/OBJC.

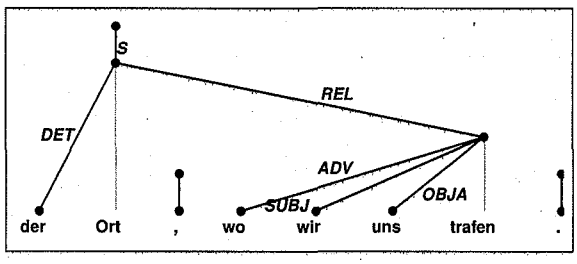
NEB oder REL?

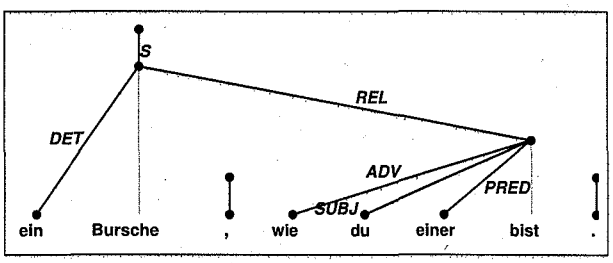
Ein Nebensatz ist REL (oder OBJC), wenn er ein Relativpronomen oder Fragepronomen in seinem Skopus hat, sonst NEB.

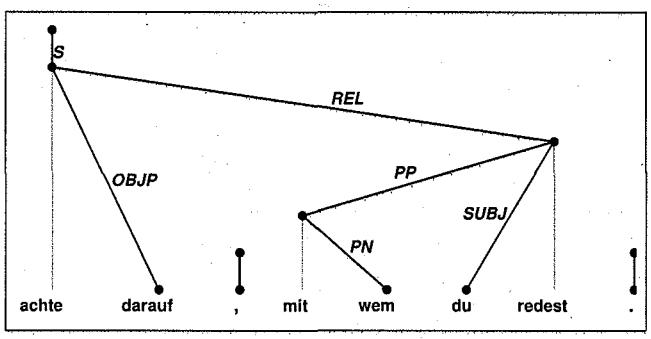




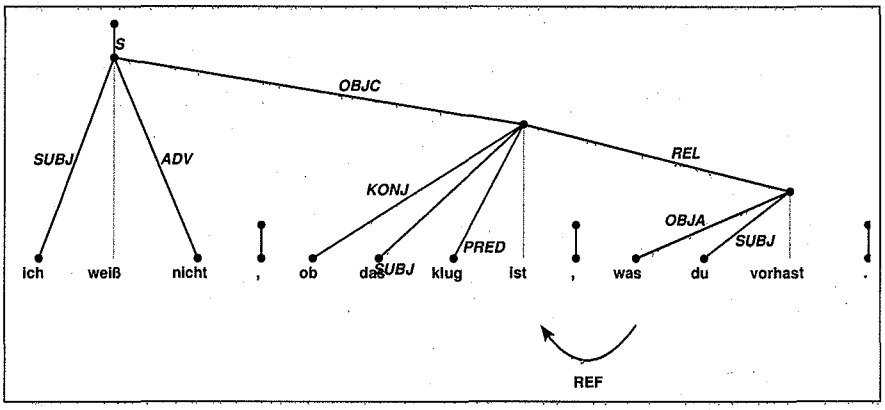






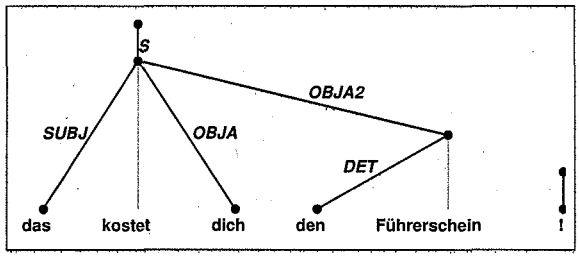


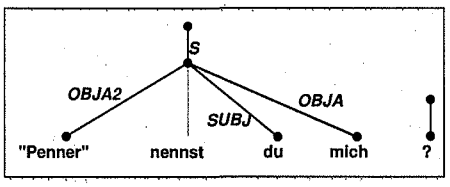
Ist ein Relativpronomen in mehrere Sätze verschachtelt, so erzwingt es das Label REL nur für den untersten, weil es nicht im Skopus der anderen Sätze steht.



OBJA oder OBJA2?

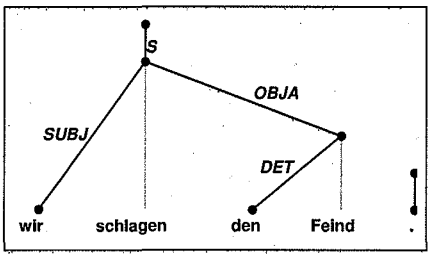
Wenn ein Verb zwei Akkusative nimmt, so ist derjenige das OBJA2, der im normalen Hauptsatz weiter rechts steht, auch dann, wenn die tatsächliche Reihenfolge umgekehrt ist.

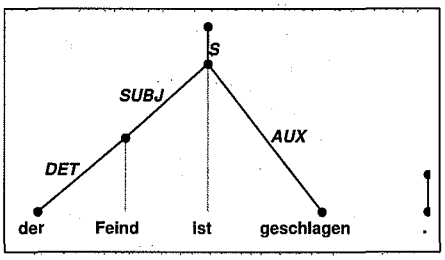


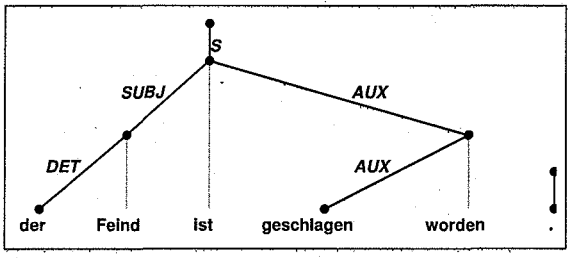


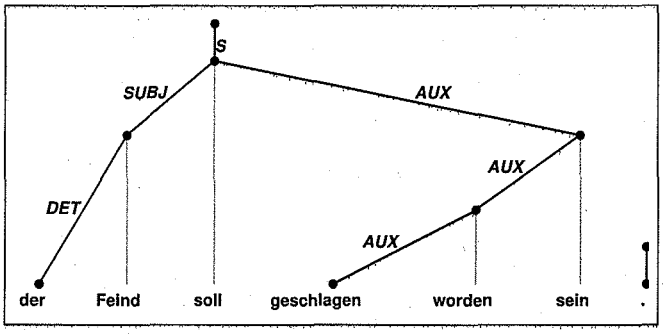
OBJA oder SUBJ?

Direkte Objekte werden im Passiv zu Oberflächensubjekten und stehen im Nominativ; daher werden sie stets als SUBJ bezeichnet. Das gilt auch dann, wenn mehrere Hilfsverben dazwischentreten.

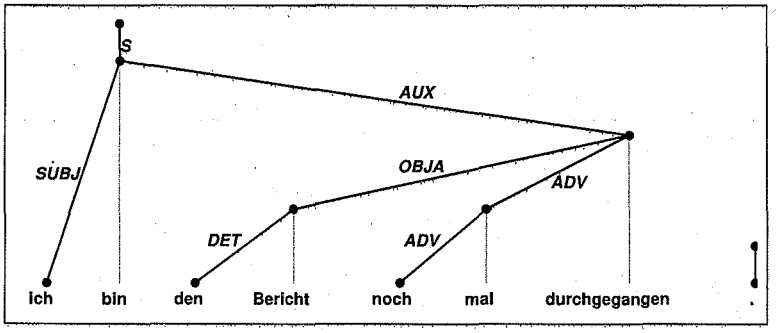




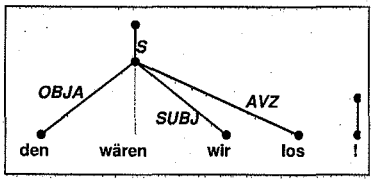




Im allgemeinen steht also niemals ein OBJA, wenn 'werden' oder 'sein' mit einem Partizip auftreten. Lediglich einige Bewegungsverben, die das Perfekt mit 'sein' bilden, tragen Akkusativobjekte:

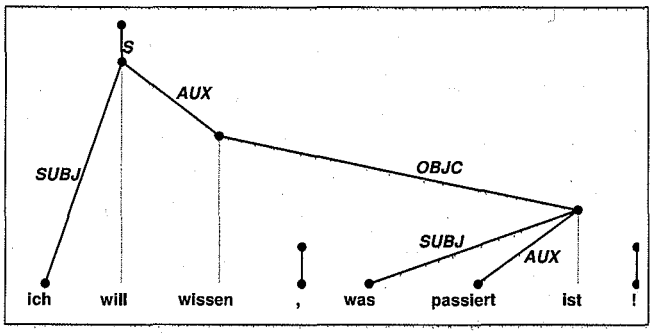


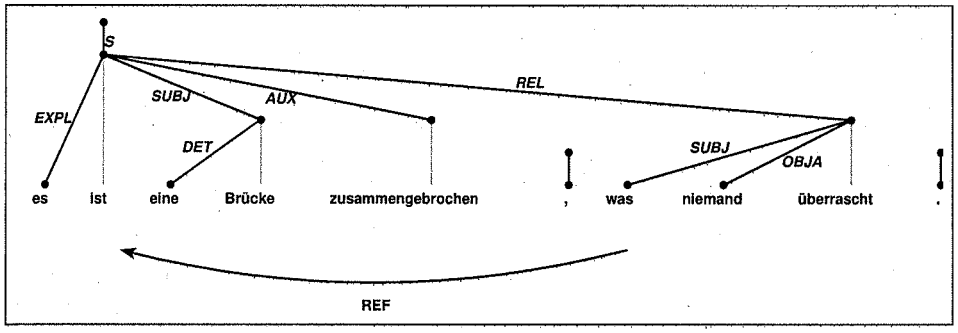
Wenn das Verb 'werden' mit Präfix auftritt, also Vollverbfunktion hat, kann jedoch sehr wohl ein Akkusativobjekt stehen:

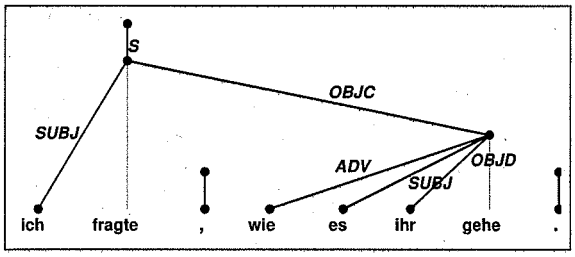


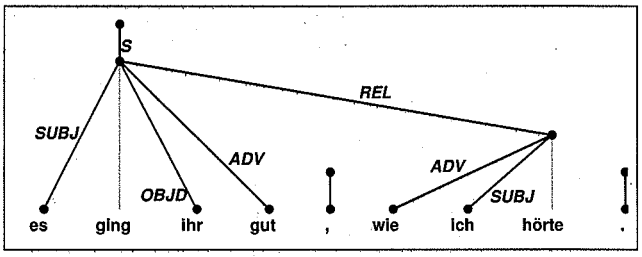
OBJC oder REL?

Untergeordnete Sätze mit Frageworten (PWS, PWAV, PWAT) sind Objektsätze, wenn der Regent es verlangt, sonst Relativsätze.





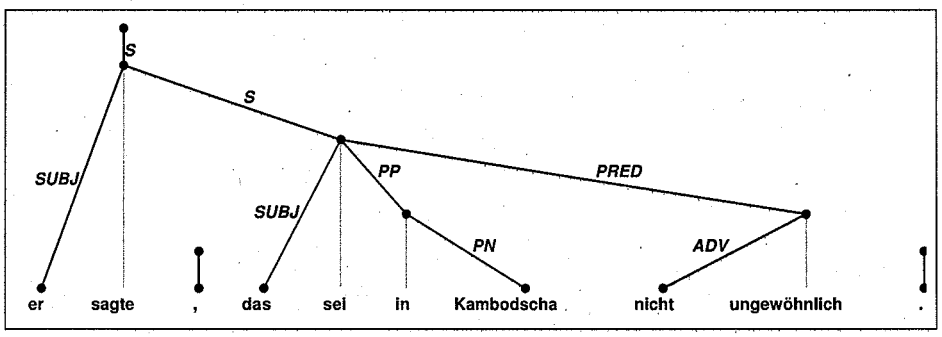


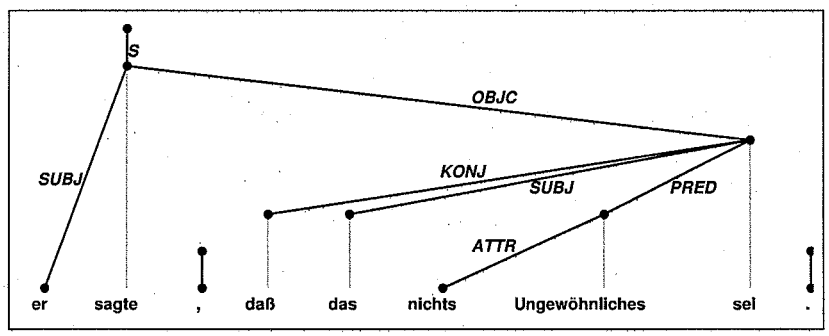


Hierbei gilt derselbe Test wie wie bei der Frage 'PRELS oder PWS?'.

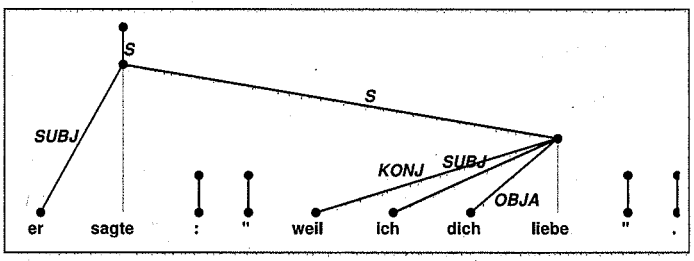
OBJC oder S?

Untergeordnete Sätze sind immer dann S, wenn sie Hauptsätze sind, also Verbzweitstellung vorliegt. Alle anderen untergeordneten Sätze sind OBJC, NEB oder REL.



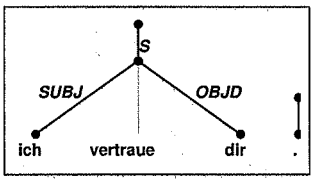


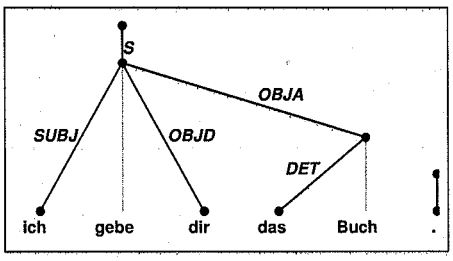
Nur wenn ein Nebensatz als direkte Rede verwendet wird, wird er als S bezeichnet.

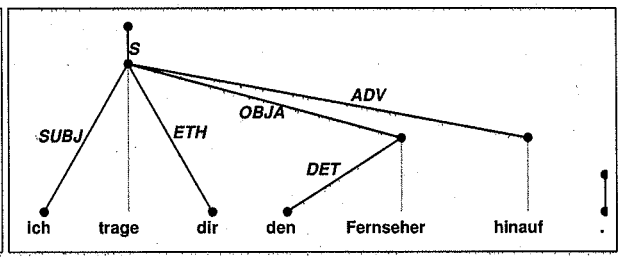


OBJD oder ETH?

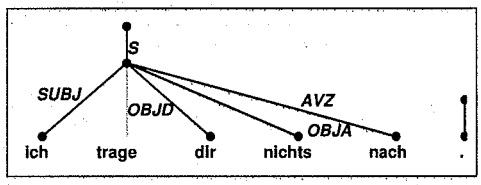
Das Label OBJD wird immer dann verwendet, wenn das Verb (evtl. in Verbindung mit einem Präfix oder bestimmten feststehenden Phrasen) regelmäßig einen Dativ nimmt. Drückt hingegen der Dativ nur die Anteilnahme, Betroffenheit oder subjektive Meinung aus, so ist ETH zu wählen.

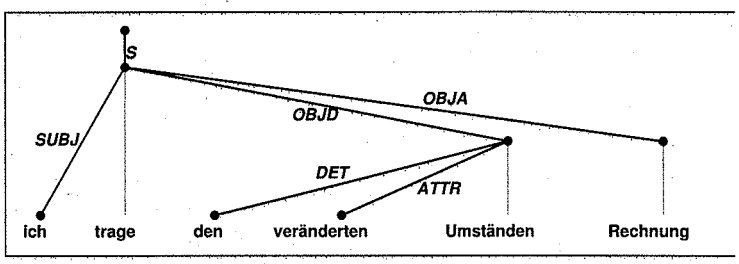






Aber:



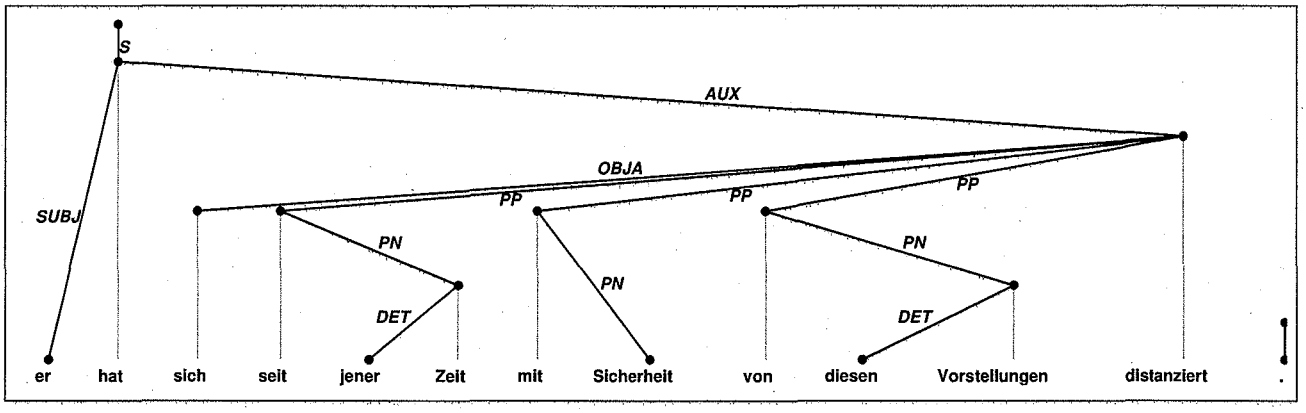


Die Abgrenzung ist oft sehr schwer zu treffen. Hat beispielsweise das Verb 'bauen' eine typische Dativ-Valenz für den Begünstigten oder nicht? Wenn ja, warum nicht auch 'basteln' oder 'erstellen'? Und wenn 'erstellen', warum nicht auch 'herstellen'? Derzeit enthält das Lexikon nicht sehr viele solche Dativ-Valenzen. Wenn man aber der Meinung ist, daß ein Verb doch ganz bestimmt indirekt oder bitransitiv ist, muß 'OBJD' gewählt, aber auch der Verbrahmen geändert werden.

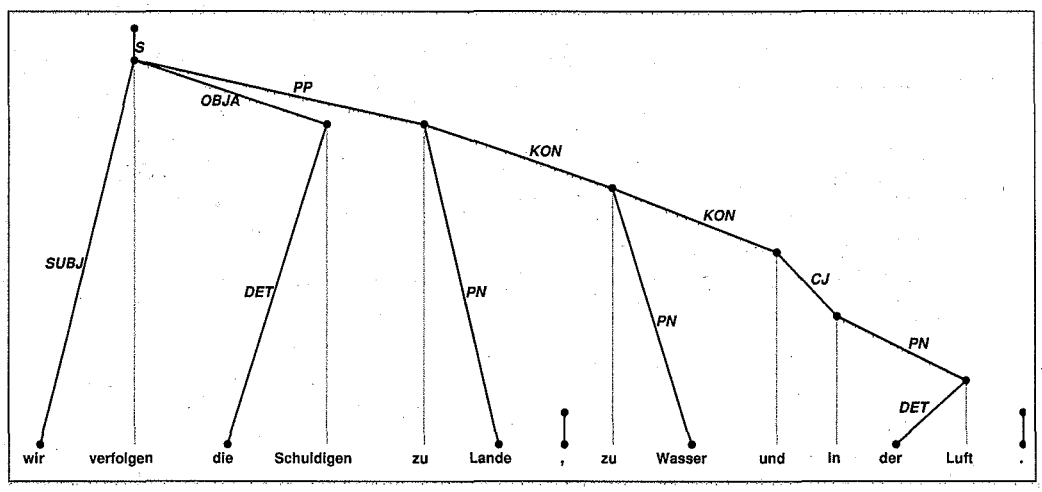
PP oder KON?

Optionale Konstituenten, die mehrfach auftreten können, sind in der Regel jedesmal mit demselben Label demselben Wort unterzuordnen. Drei Präpositionalphrasen im selben Satz-

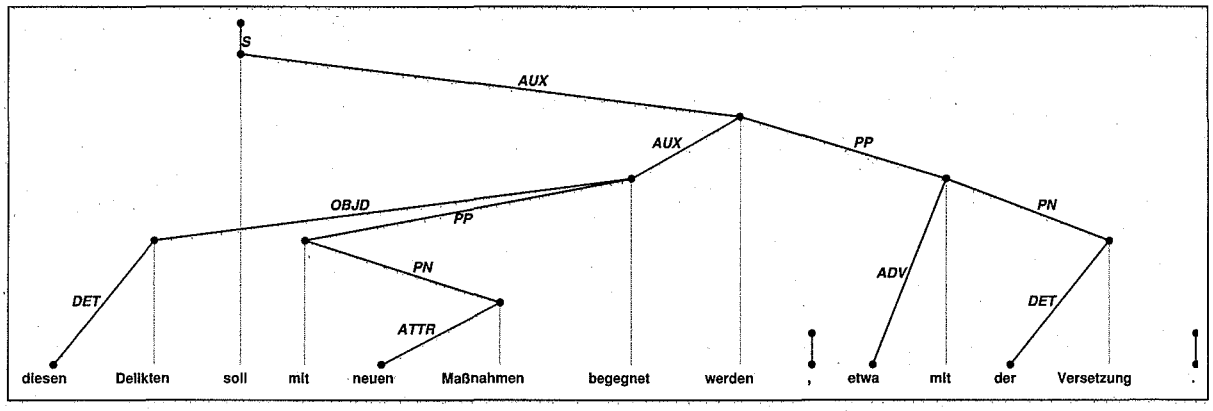
feld sind also jeweils PP.



Nur wenn tatsächlich eine Konjunktion auftritt, sollten KON gewählt werden.



Wenn zwei Konstituenten in verschiedenen Satzfeldern stehen, also nur nichtprojektiv verbunden werden können, sollten sie stets parallel untergeordnet werden, auch wenn sie inhaltlich sehr eng zusammengehören.



PP oder OBJP?

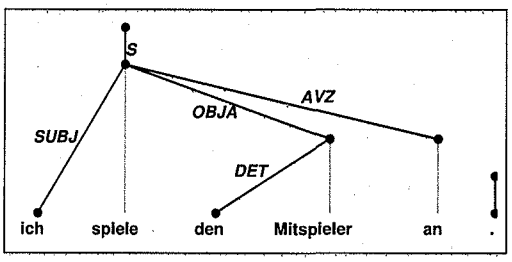
Präpositionalobjekte sind solche PP, die nicht frei zu einem Verb treten, sondern mit ihm zusammen eine nicht kompositionale Bedeutung erzeugen. (TIGER-Annotationsrichtlinien: "Präpositionalobjekte sind Präpositionalphrasen, die infolge eines Abstraktionsprozesses an das Verb gebunden sind.")

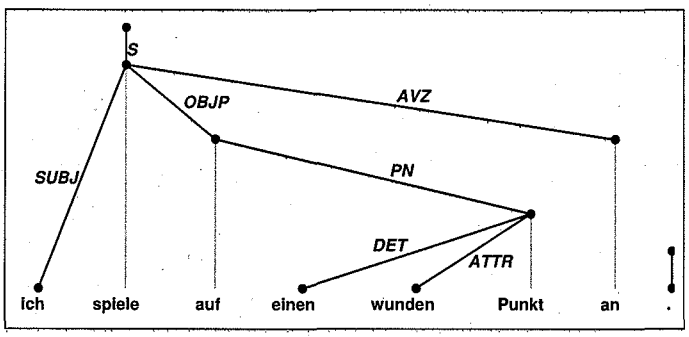
Für die Unterscheidung zwischen OBJP und PP gibt es verschiedene Anhaltspunkte. Typischerweise treten diejenigen Präpositionen, die eine OBJP-Beziehung zu einem Verb eingehen können, auch öfter zusammen mit diesem Verb auf, als der Zufall rechtfertigen würde. So ist etwa die Kombination 'beruhen auf' auch eine starke Kollokation, wesentlich stärker als etwa 'beruhen an', obwohl beide Präpositionen etwa gleich häufig sind.

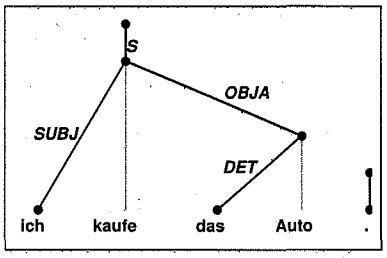
Wichtiger als bloße Häufigkeit ist jedoch das schwerer erfaßbare semantische Kriterium. In einem Präpositionalobjekt ist typischerweise die ursprüngliche Bedeutung der Präposition bis zur Bedeutungslosigkeit verblaßt: z.B. hat 'an' in 'glauben an' keine Ortsbedeutung, und 'beruhen' impliziert keine *räumliche*, sondern kausale Beziehung zwischen Subjekt und Präpositionalobjekt.

Dieser Test ist aber nur einseitig anwendbar, denn auch in vielen freien PP ist die Bedeutung verblaßt; so hat z.B. das 'im' in 'im Grunde' auch keine Ortsbedeutung mehr.

Ein rein syntaktisches Kriterium ist, daß ein Präpositionalobjekt bei einem eigentlich transitiven Verb das Akkusativobjekt vertreten kann:



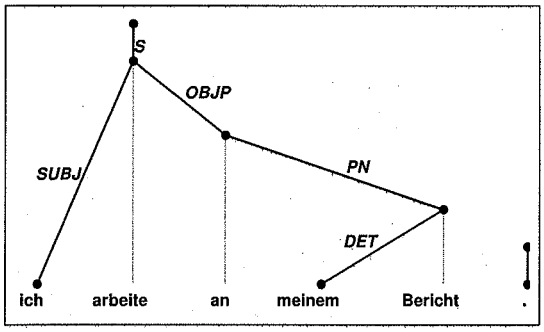


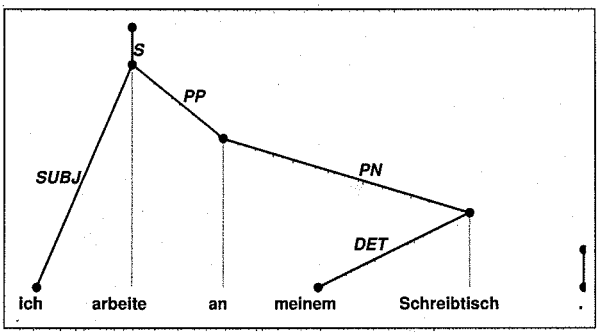


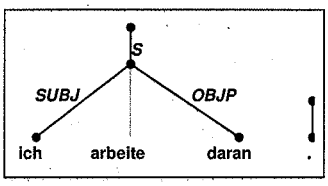
- *Ich kaufe auf/OBJP der Straße.

Dieser Test funktioniert nur in einer Richtung, denn viele Präpositionalobjekte treten mit ansonsten intransitiven Verben auf (z.B. 'zurückgreifen'+'auf').

Wenn die Ersetzung der PP durch ein Pronominaladverb zu einer Sinnänderung führt, handelt es sich mit Sicherheit um eine normale PP:

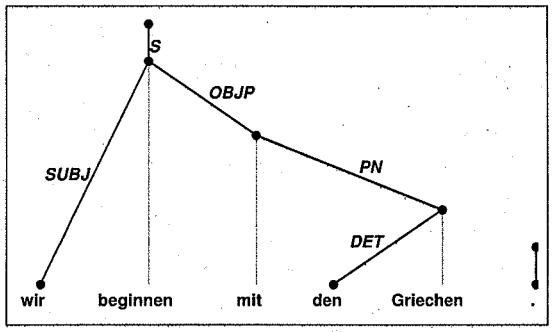




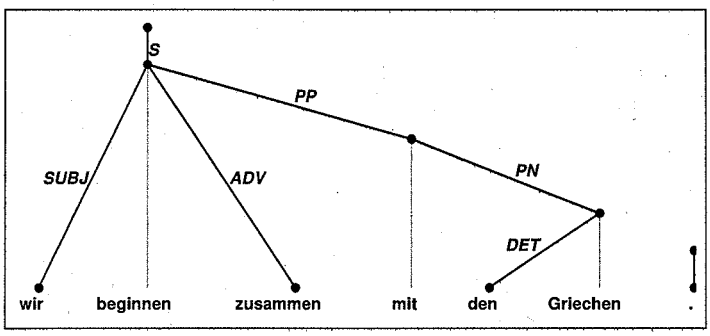


(=an dem Bericht, nicht dem Schreibtisch)

Wenn die adverbiale Erweiterung der PP zu einer Sinnänderung führt, handelt es sich um ein Präpositionalobjekt:

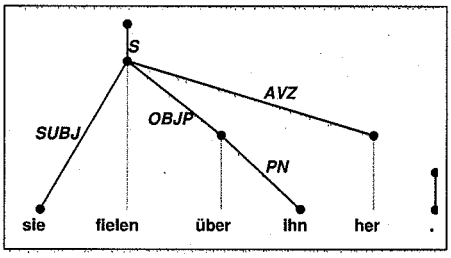


(= behandeln als erstes die Griechen)



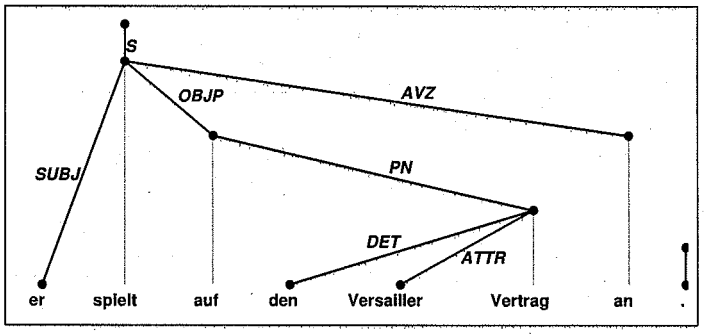
(= fangen gleichzeitig mit ihnen an)

Wenn ein Verb ohne seine Präposition ungrammatisch wird, handelt es sich sicher um ein Präpositionalobjekt:

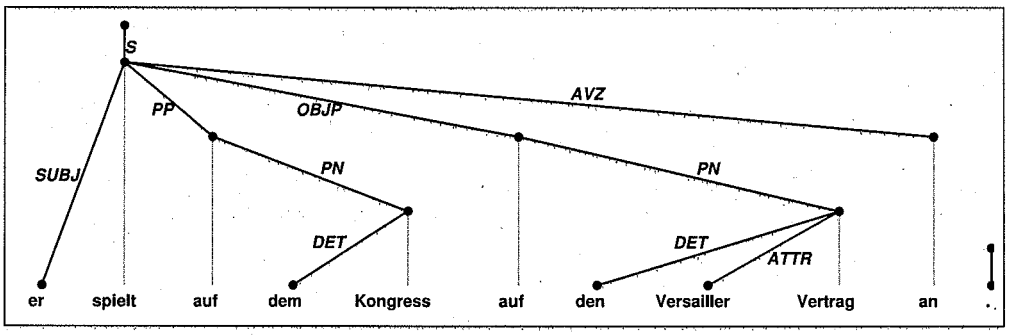


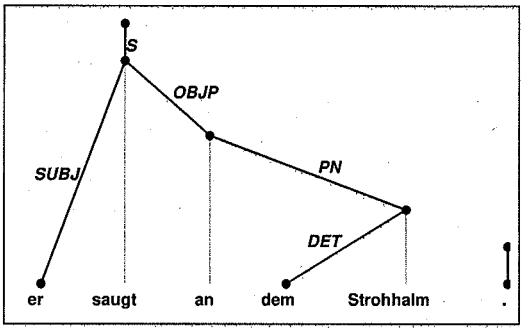
- *Sie fielen her.

Wenn ein Verb eindeutig Richtungs- oder Ortsbedeutung hat, muß ein Präpositionalobjekt derselben Lokationsklasse angehören, während freie PP dieser Beschränkung nicht unterliegen:

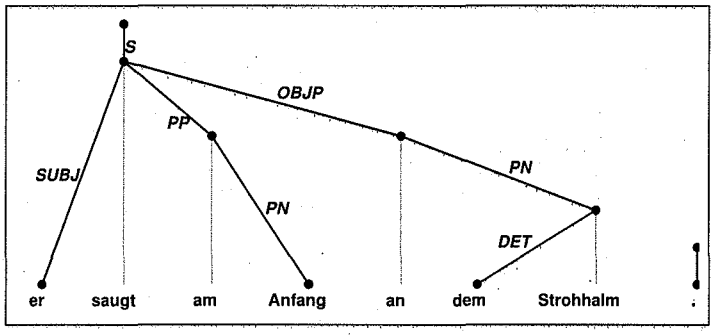


- *Er spielt auf/OBJP dem Versailler Vertrag an.

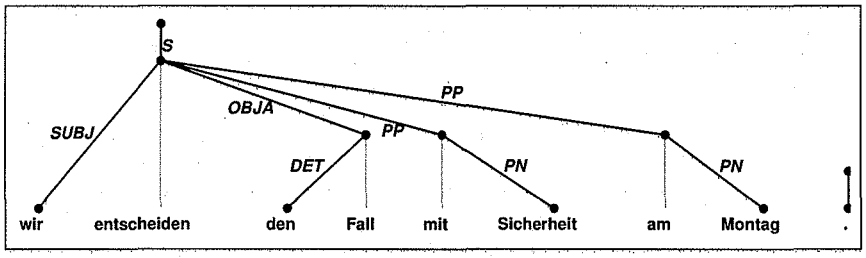




- *Er saugt an/OBJP den Strohhalm.

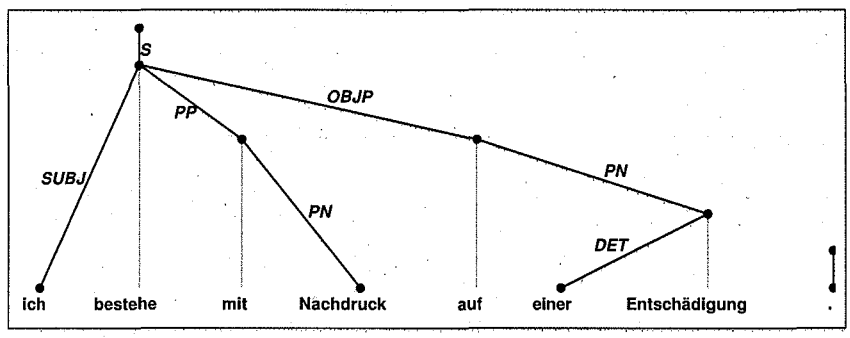


Präpositionalobjekte müssen eindeutig sein, während freie PP mehrfach auftreten können.

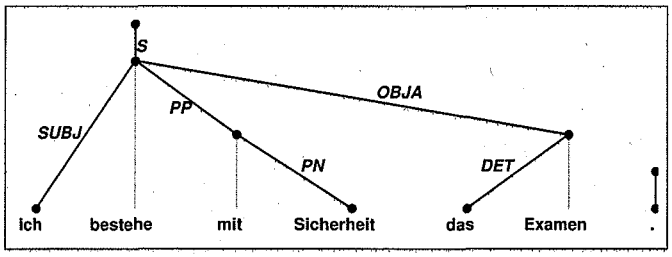


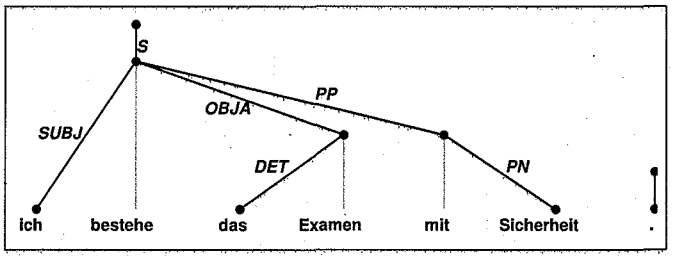
- *Wir entscheiden über/OBJP den Fall über/OBJP andere Fragen.

Präpositionalobjekte stehen nach anderen adverbialen Bestimmungen, während normale PP dieser Beschränkung nicht unterliegen.

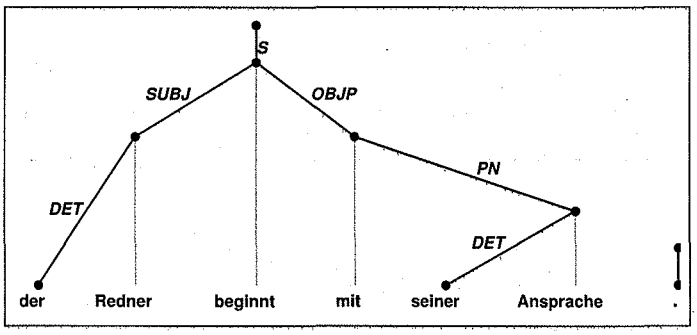


- *Ich bestehe auf einer Entschädigung mit Nachdruck.

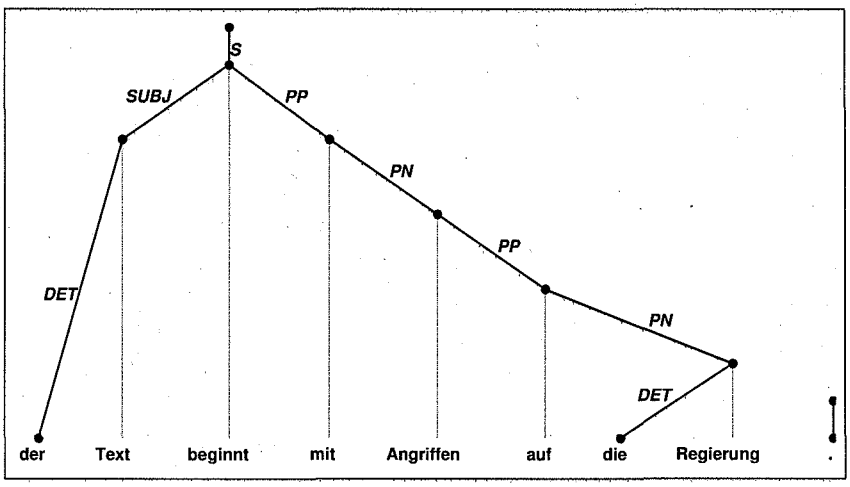




Es gibt Verben, die sowohl eine kompositionale als auch eine nicht-kompositionale Präferenz für dieselbe Präposition haben.

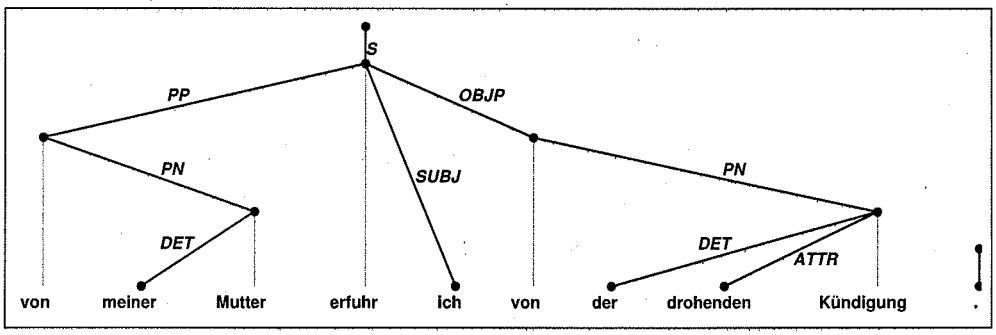


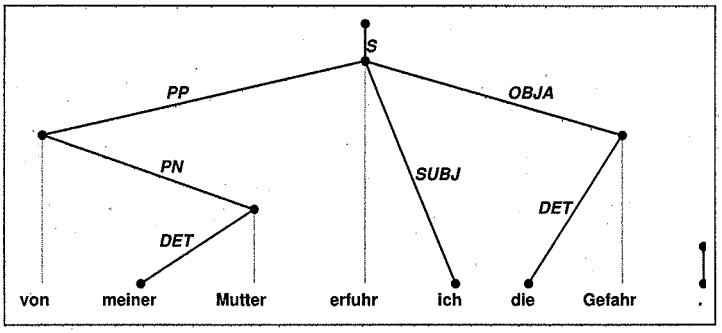
(= fängt an zu sprechen)



(≠ fängt an, die Regierung anzugreifen)

Beide Möglichkeiten können sogar gleichzeitig auftreten; in diesem Fall kann nur eines der beiden Auftreten OBJP sein. Es hilft oft der Versuch, eine der beiden PP durch das entsprechende Objekt zu ersetzen:

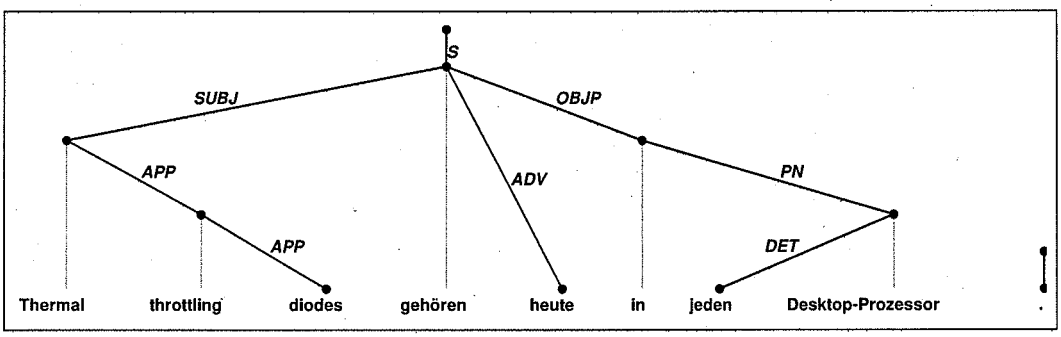




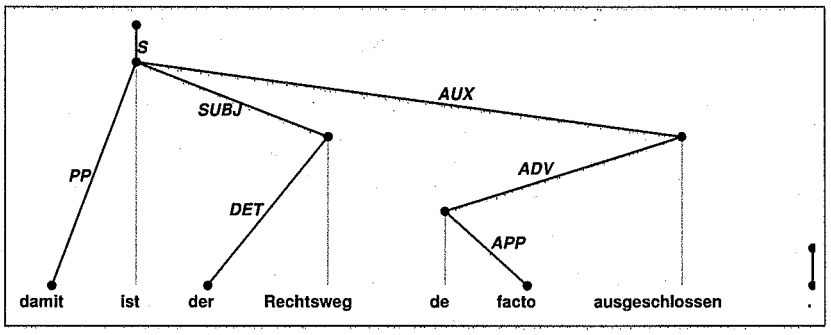
- *Von/PP der drohenden Kündigung erfuhr ich meine Mutter/OBJA.

S oder ADV?

Fremdsprachliches Material ist als APP, SUBJ etc. zu annotieren, wenn es die Rolle einer normalen NP ausfüllt.

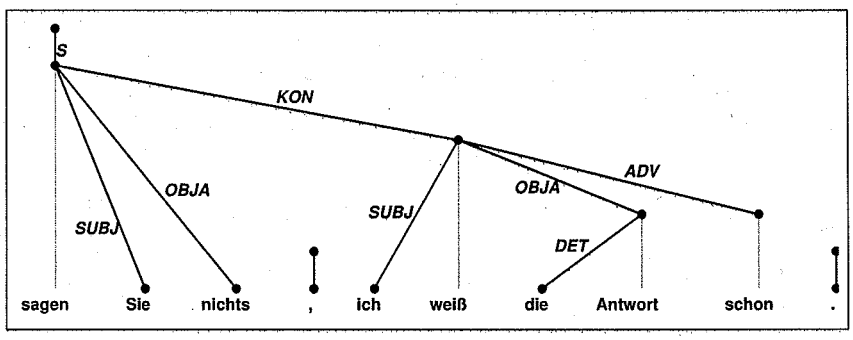


Nur wenn eine ganze fremdsprachliche Phrase auch im Deutschen eindeutig adverbiale Bedeutung besitzt und auch so benutzt wird, ist ADV zu wählen.

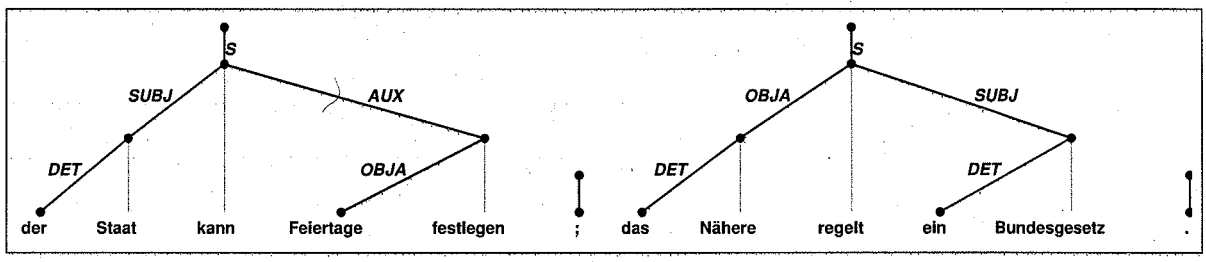


S oder KON?

Stehen zwei ganze Sätze ohne Konjunktion nebeneinander, so wird der zweite dem ersten durch KON untergeordnet.

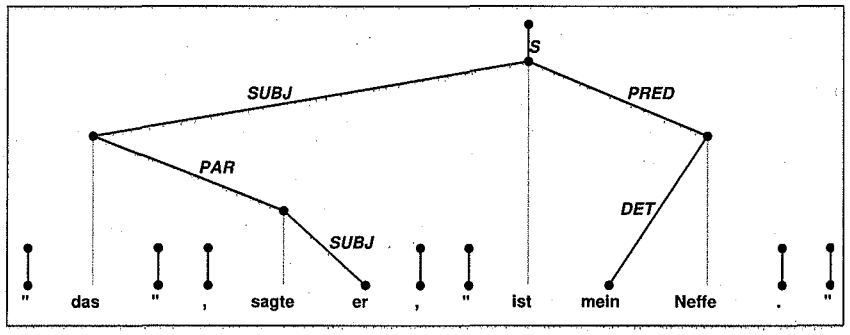


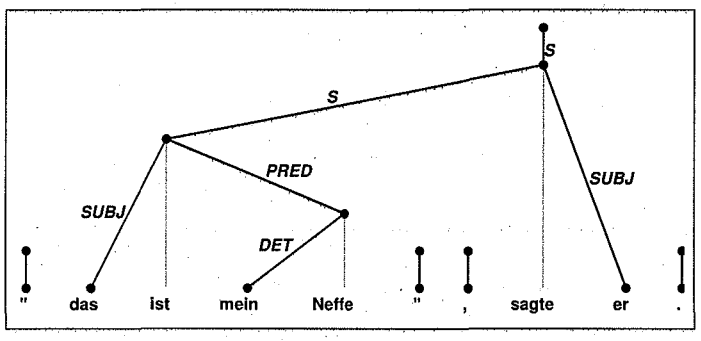
Wenn die Sätze durch : oder ; getrennt sind, werden stattdessen zwei Satzwurzeln annotiert.



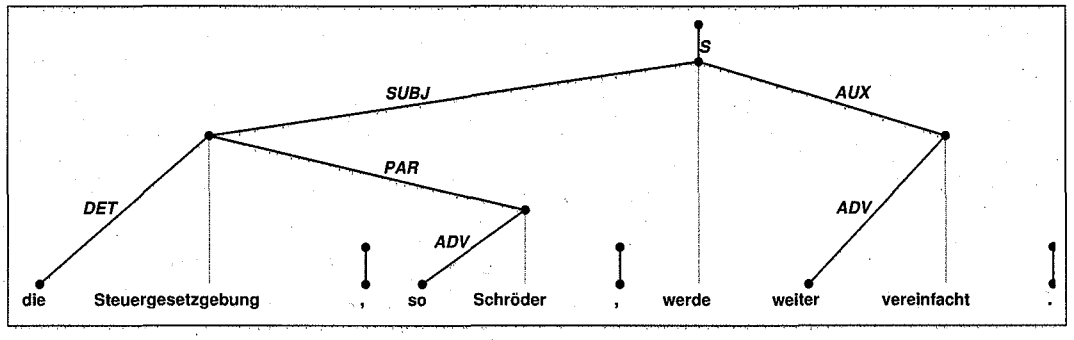
S oder PAR?

Das Label PAR wir nur verwendet, wenn die normale Unterordnung durch S einen nichtprojektiven Baum erzeugen würde. PAR wird also verwendet, wenn der Matrixsatz zwischen zwei Teilen des untergeordneten Satzes steht.

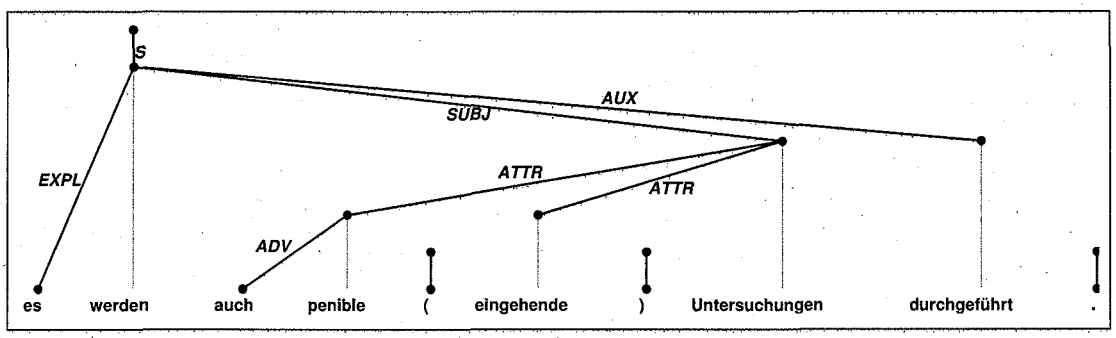


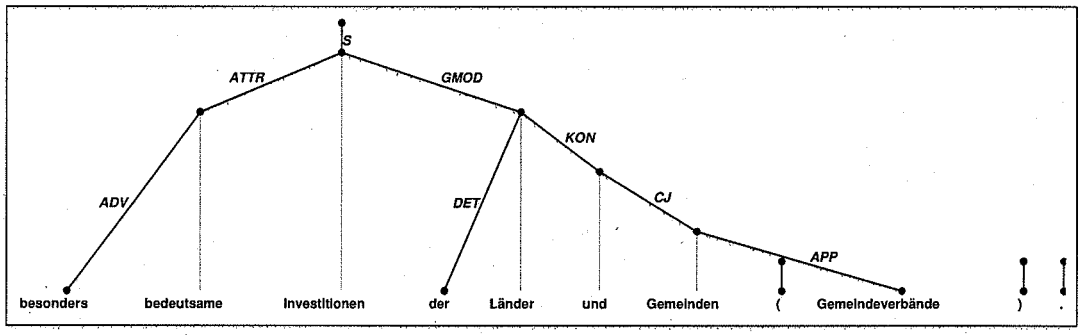


Wenn das Matrixverb ausgelassen wird, wird es oft durch 'so' ersetzt. In diesem Fall wird das logische Subjekt als PAR annotiert:

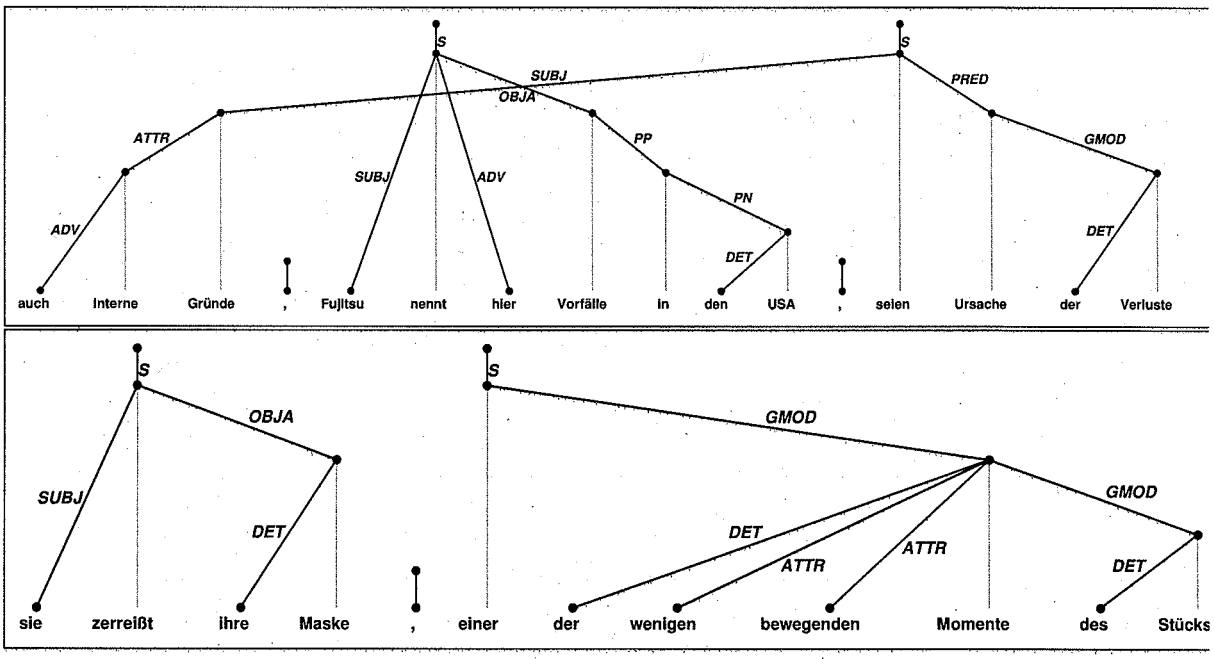


PAR wird nicht für andere Einschübe verwendet. Überzählige oder durch Komma markierte Konstituenten erfüllen entweder dieselbe Funktion wie die primäre Konstituente, oder sie sind dieser als APP oder KON beigeordnet.



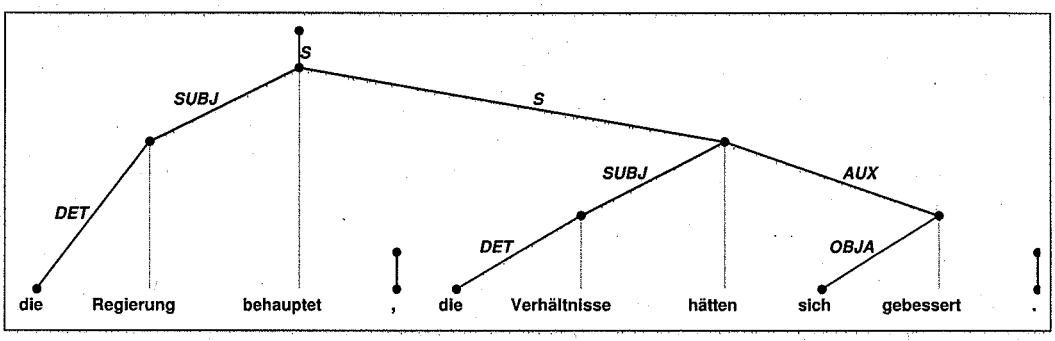


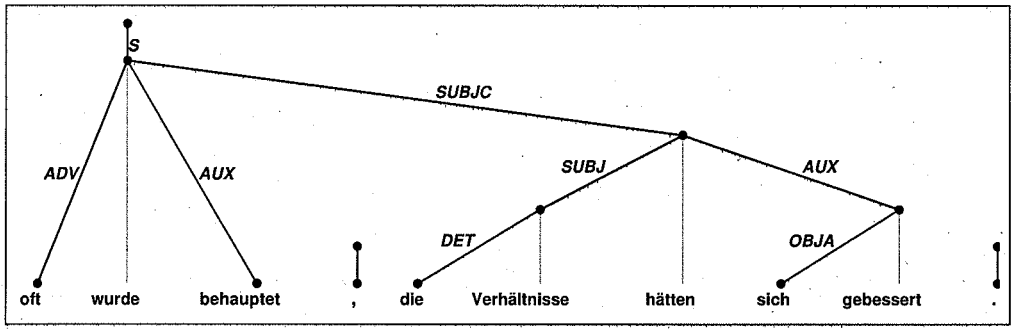
Einschübe, in denen ein VP durch eine NP erläutert wird oder umgekehrt, können überhaupt nicht repräsentiert werden und sind bilden ein zusätzliches S.



S oder SUBJC?

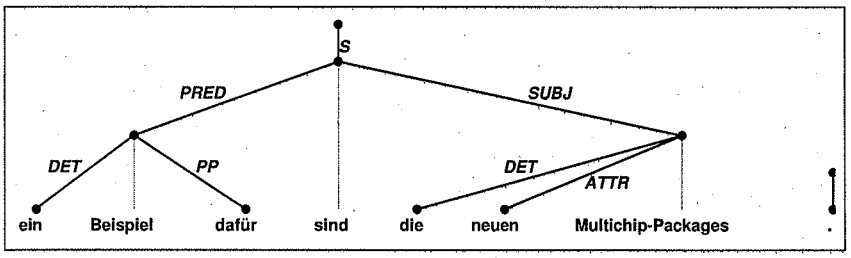
Der Objektsatz des Hauptverbs wird im Passiv zum Subjektsatz, ist also immer SUBJC und nicht S.



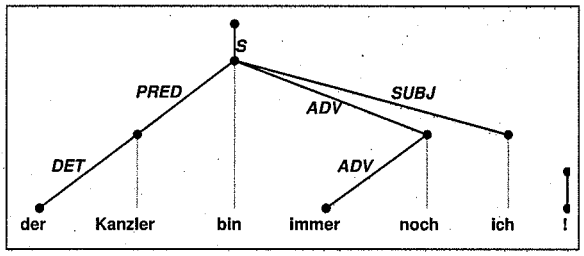


SUBJ oder PRED?

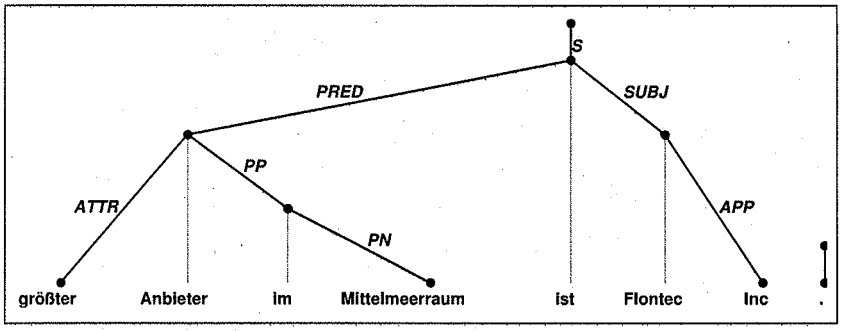
In vielen Sätzen können Subjekt und Prädikat miteinander verwechselt werden. Da sowohl Subjekt als auch Prädikat im Nominativ stehen, läßt sich kein Kasustest durchführen. Auch die Reihenfolge der beiden NP ist im allgemeinen *nicht* ausschlaggebend für die Unterscheidung. Möglich ist allerdings der Test auf Kongruenz in Bezug auf Person und Numerus:

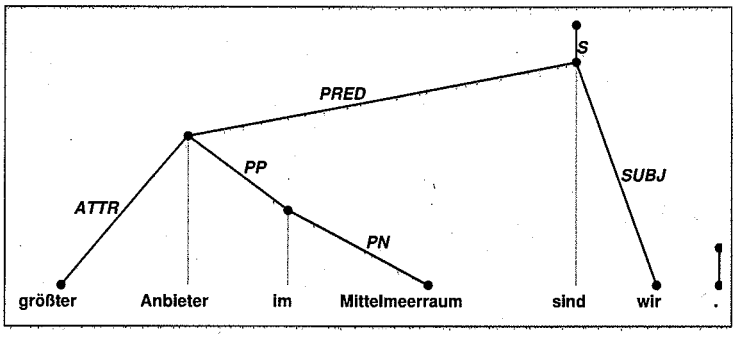


Da das Verb im Plural steht, kann 'Beispiel' nicht das Subjekt sein, sondern nur das Prädikat.



Da das Verb in der ersten Person steht, muß 'ich' das Subjekt sein. Da fast alle Verben in der dritten Person auftreten, ist dieser Test selten anwendbar; man kann aber versuchen, den Satz sinnerhaltend aus der dritten in die erste Person umzuformulieren.

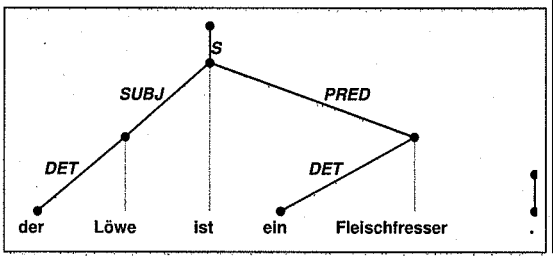


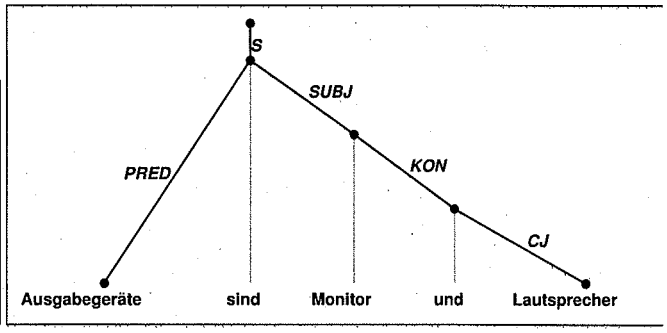


- *Wir sind Flontec Inc.

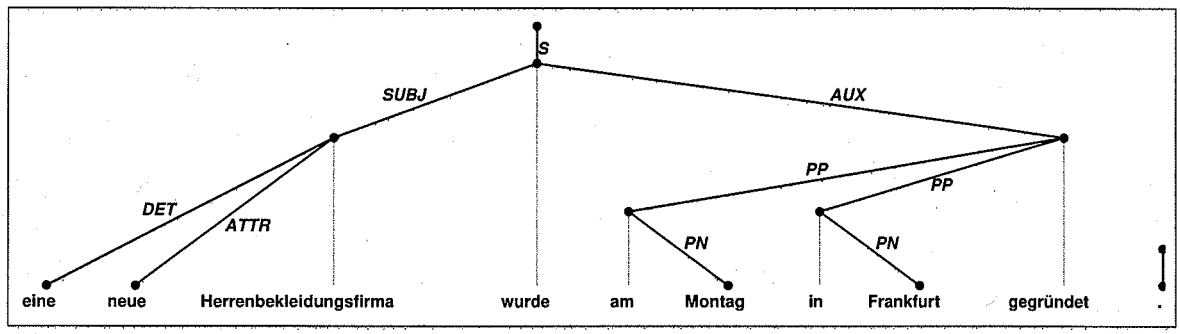
Die korrekte Paraphrase von (1) (geäußert von einem Flontec-Manager) ist (2) und nicht (3); also ist in (1) 'Flontec' Subjekt und nicht Prädikat.

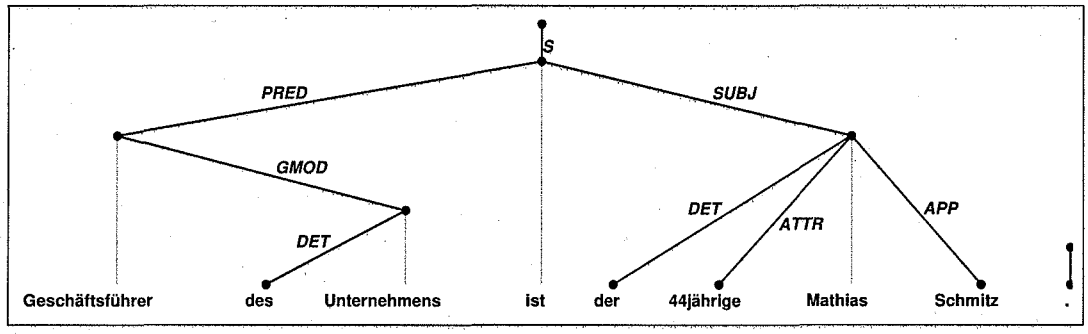
Wenn ein Satz ausdrückt, daß ein Mensch oder Gegenstand eine bestimmte Rolle ausfüllt, so die Rolle PRED und ihr Füller SUBJ:





In solchen Fällen ist oft (aber nicht immer) das Prädikat eine allgemeinere NP als das Subjekt. Auch steht das Prädikat tendentiell mir dem unbestimmten und das Subjekt mit dem bestimmten Artikel. Diese Faustregeln sind aber fehlbar. Wichtiger ist die genaue Betrachtung der inhaltlichen Aussage:



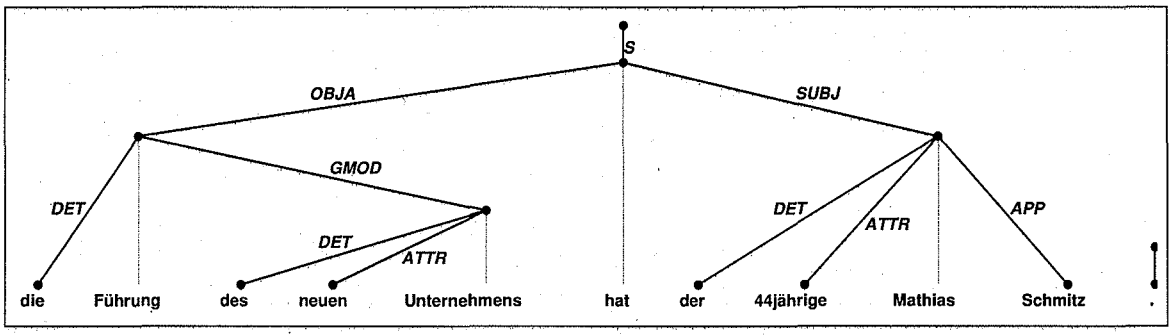


Hier ist offensichtlich gemeint, daß Herr Schmitz in dem Unternehmen die Rolle des Geschäftsführers ausfüllt.

- Das hob Rainer Zorbach, Leiter des Bereichs E-Business der HypoVereinsbank, bei der Einweihung der neuen Räumlichkeiten des TC Trust Centers in Hamburg hervor. Das TC Trust Center/SUBJ ist eine gemeinsame Tochter/PRED von Deutscher Bank, Dresdner Bank, Commerzbank und HypoVereinsbank und bietet Lösungen für sichere Transaktionen im Internet an.

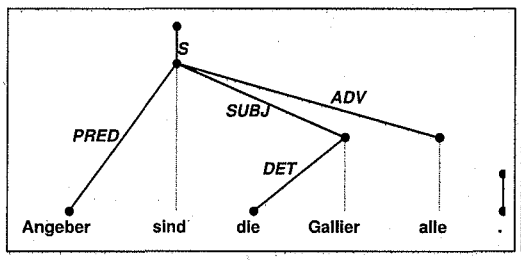
Hier würde die Belegung Center/PRED bedeuten, daß jedes Unternehmen die Position 'TC Trust Center' vergibt und sie in diesem Fall durch eine Tochter der vier Banken gefüllt wird. Das aber ist nicht der Fall, sondern 'TC Trust Center' ist offenbar der Name einer Firma, deren Rechtsstellung im zweiten Satz genauer erläutert wird.

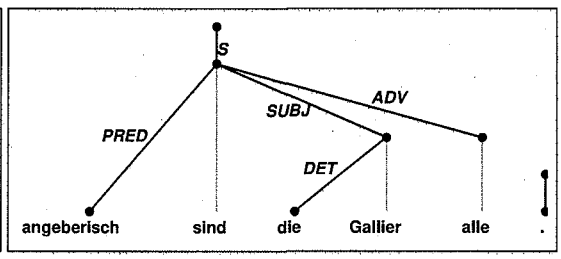
Auch hier kann eine Umformulierung bei der Entscheidung behilflich sein. Läßt sich eine der beiden NP zusammen mit dem Verb sinnerhaltend in eine andere VP umformulieren, so ist diese meist das Prädikat:



- *Geschäftsführer des Unternehmens ist 44 Jahre alt und hört auf den Namen Mathias Schmitz.

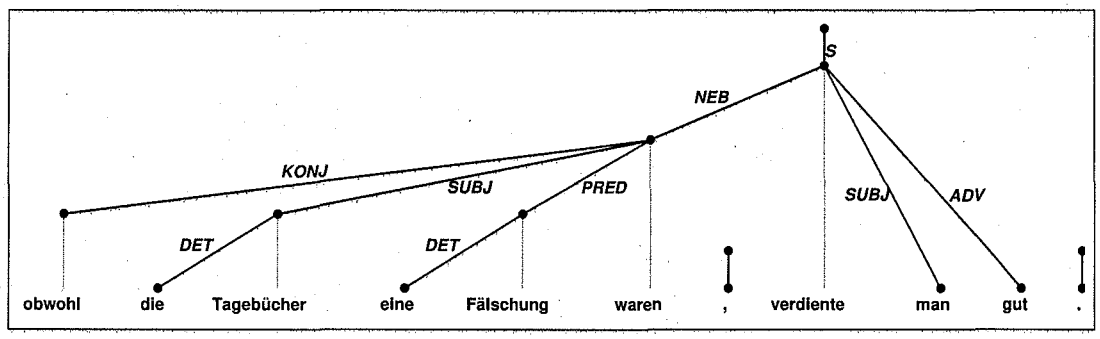
Auch wenn sich eine von zwei NP zu einem Adjektiv umformulieren läßt, so ist diese das Prädikat.





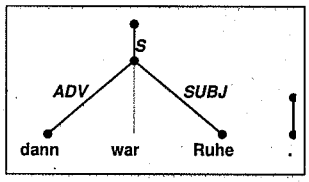
- *Angeber sind gallisch alle.

Stehen beide NP im Mittelfeld, so ist fast immer die erste das Subjekt.

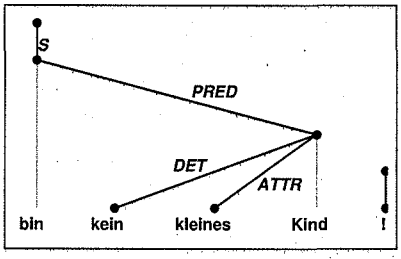


- *Obwohl eine Fälschung die Tagebücher waren, verdiente man gut.

Wenn nur eine NP vorhanden ist, ist diese normalerweise das Subjekt.

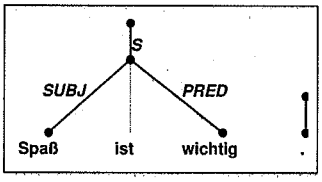


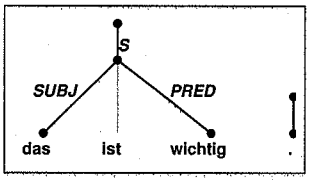
Nur wenn das Subjekt offensichtlich ausgelassen worden ist, ist die verbleibende NP PRED.

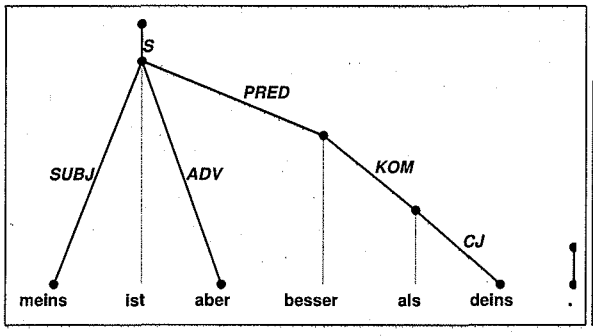


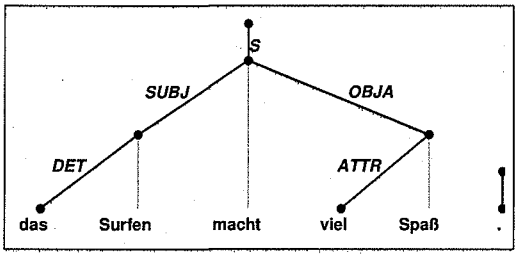
SUBJ oder SUBJC?

Nominalphrasen jeder Art sind SUBJ, wenn sie Subjektbedeutung haben. Das gilt auch für substantivierte Infinitive, die der Kategorie NN angehören.

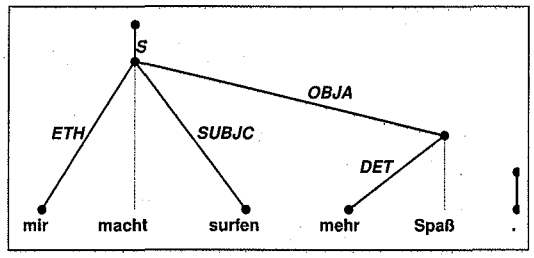


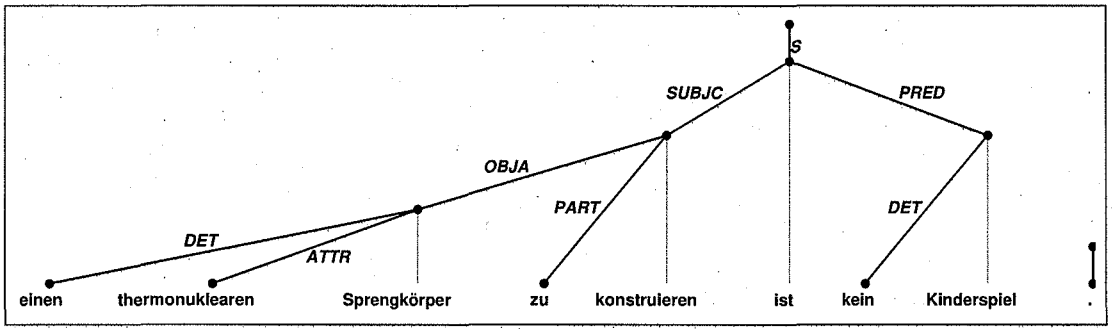






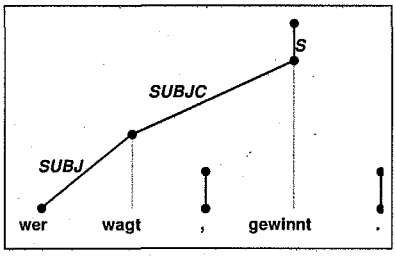
SUBJC wird nur für echte Verben in Subjektposition verwendet.



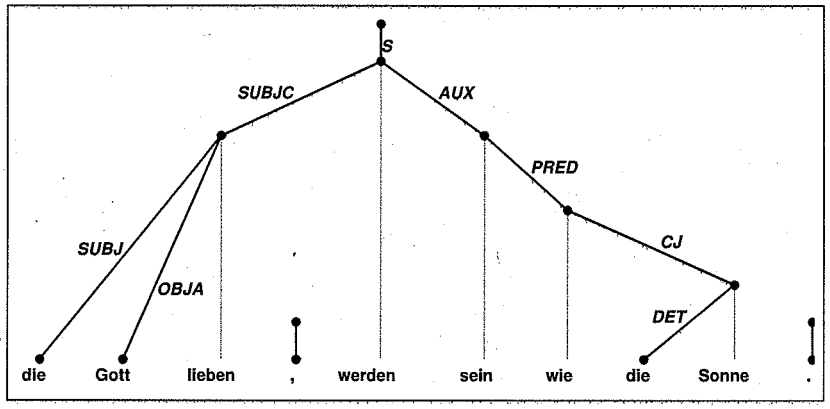


SUBJC oder REL?

Relativsätze können das Subjekt vertreten, sowohl solche mit 'wer' etc. als auch solche mit 'der' etc. Diese Anbindung ist immer zu wählen, wenn der Hauptsatz kein eigenes Subjekt hat und eine Paraphrase durch 'die, welche oder 'derjenige, der' möglich ist.



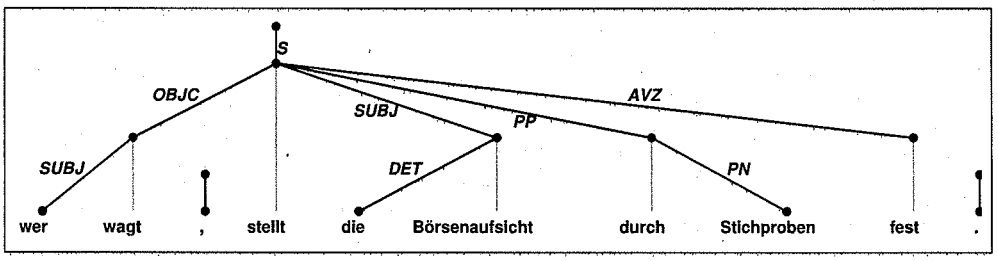
(= derjenige, der wagt)



(= die, welche Gott lieben)

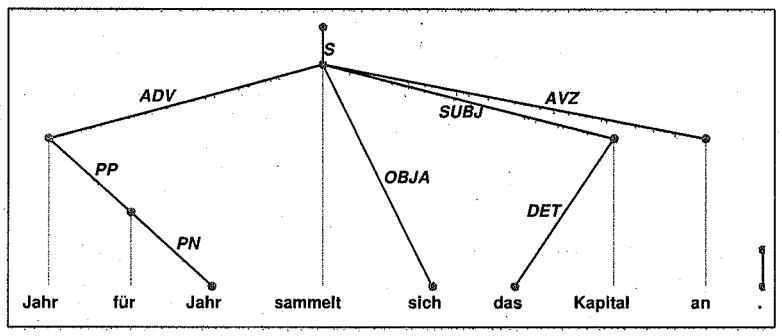
Anderenfalls ist REL oder OBJC zu wählen.

- Wer wagt/REL, der gewinnt.



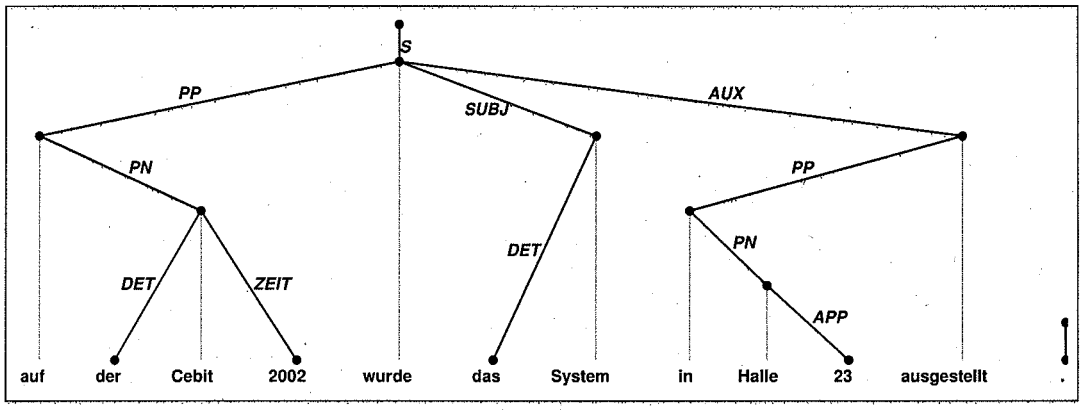
ZEIT oder ADV?

Wenn eine absolute Zeitangabe auch als Reduplikation aufgefaßt werden könnte, so ist ADV anzunehmen und nicht ZEIT.



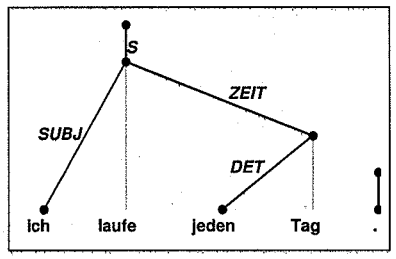
ZEIT oder APP?

Wenn Zahlen als Bestandteil von mehrteiligen NP verwendet werden, sind sie normalerweise als APP zu bezeichnen. Handelt es sich jedoch um eine Jahreszahl, die ausdrückt, daß eine Veranstaltung tatsächlich in diesem Jahr stattfand, ist ZEIT zu verwenden.

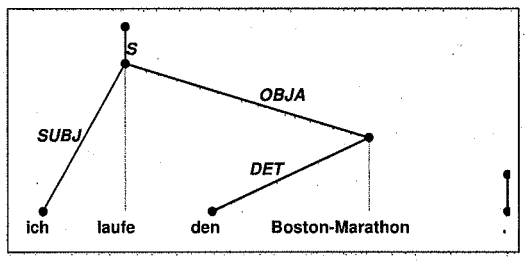


ZEIT oder OBJA?

Die Funktion ZEIT ist nur für Zeitausdrücke erlaubt. Entscheidend ist, ob der Satz durch adverbiale Zeitbestimmung paraphrasiert werden kann oder nicht.

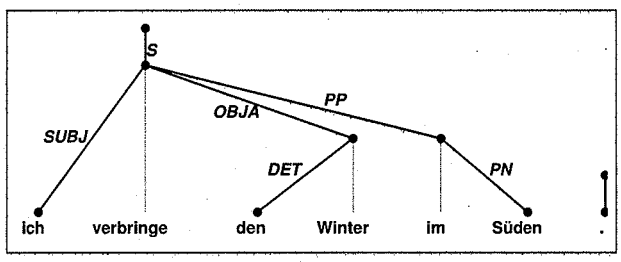


(= an jedem Tag, also temporal)



(= beim Boston-Marathon, also lokal)

Wenn beides möglich ist, ist dennoch OBJA zu wählen, wenn das Objekt sonst fehlen würde:



1.2.5 Welche Struktur?

Im allgemeinen soll der Syntaxbaum stets so gezeichnet werden, daß er die Prinzipien von Kongruenz, Topologie, Wortstellung etc. formal erfüllt, also die logisch richtige *Oberflächenstruktur* aufweist. Es kann in verschiedenen Fällen zu Konflikten der Oberflächenstruktur mit der offensichtlich beabsichtigten Bedeutung kommen; hier zwei Beispiele in einem Satz:

- Und es geschah in den Tagen Amrafels, des Königs von Schinar, Arjochs, des Königs von Ellasar, Kedor-Laomers, des Königs von Elam, und Tidals, des Königs von Gojim, dass sie Krieg führten mit Bera, dem König von Sodom, und mit Birscha, dem König von Gomorra, Schinab, dem König von Adma, und Schemeber, dem König von Zebojim, und mit dem König von Bela, das ist Zoar.

Die Kriegsparteien stehen also in folgender Beziehung zueinander:

Wir	Sie
Amrafel (Schinar)	Bera (Sodom)
Arjoch (Ellasar)	Birscha (Gomorra)
Kedor-Laomer (Elam)	Schinab (Adma)
Tidal (Gojim)	Schemeber (Zebojim)
	Zoar (Bela)

Inhaltlich gehören also Bera, Birscha, Schinab, Schemeber und Zoar zueinander; aber die fünf Könige sind durch drei Vorkommen von 'mit' und drei Vorkommen von 'und' unregelmäßig miteinander verbunden. Die Koordination muß so gezeichnet werden, daß jeweils Nomen mit Nomen und Präposition mit Präposition beigeordnet wird. Die Syntaxstruktur hat demnach die Form "mit (...), und mit (...), und mit (...)." Sie hat also drei Teile statt fünf, wobei der mittlere selbst eine Koordination aus drei Nominalphrasen ist.

Des weiteren ist die Ortszugehörigkeit auf verschiedene Weise ausgedrückt. Bei den ersten acht Königes wird sie mit 'von' konstruiert ("Amrafels, des Königs von Schinar"), so daß die inhaltlich zusammengehörigen Begriffe mit `APP` und `PP` direkt verbunden werden können. Der neunte aber wird durch einen Hauptsatz erläutert ("...dem König von Bela, das ist Zoar."), der nicht mit einer Nominalphrase beigeordnet werden kann. Stattdessen muß 'ist' das entfernte 'geschah' modifizieren, obwohl es inhaltlich viel enger zum letzten 'König' gehört.

Manchmal geht die inhaltliche Verwirrung so weit, daß sowohl die sinnentsprechende als auch die logisch richtige Struktur wichtige Regeln verletzen:

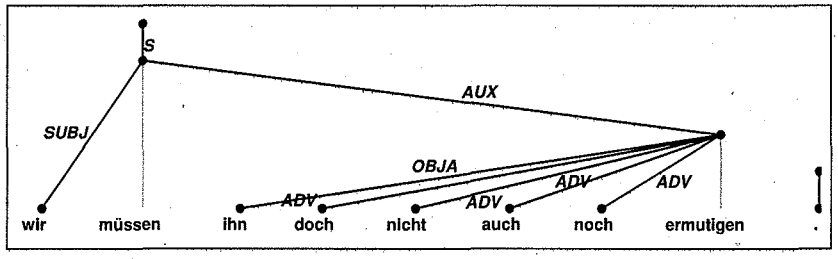
- Weiter heißt es bei NEC, dass die weltweit starke Nachfrage nach PCs und allgemein nach Halbleitern das positive Ergebnis ermöglichten.

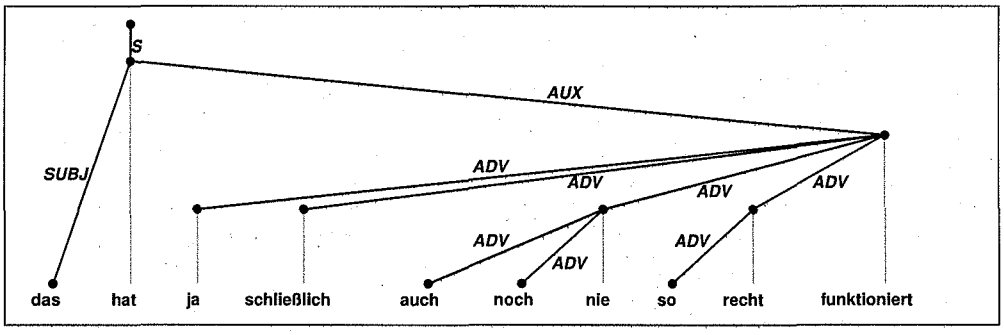
Entweder wird 'Nachfrage nach PCS' mit 'nach Halbleitern' zu einer Plural-NP verbunden, was aber den Kategorieregeln widerspricht; oder 'nach' wird logisch richtig an 'nach' koordiniert, wodurch aber die Numerus-Kongruenz zum Verb verletzt wird.

Anbindung von ADV

Aneinander (seriell) oder an dasselbe Wort (parallel)?

Adverben (ADV und ADJD) treten oft in Gruppen auf:

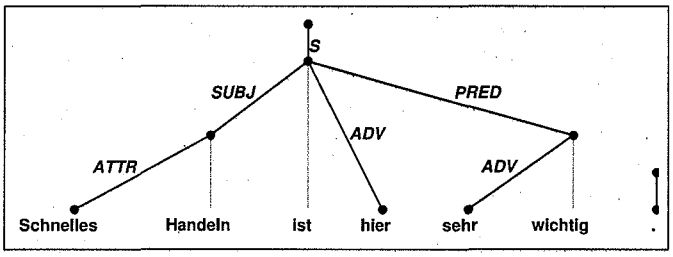




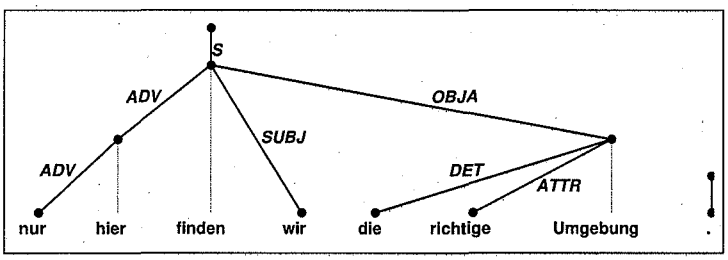
Im allgemeinen modifizieren in solchen Adverbgruppen alle Worte dasselbe Wort (meist das Vollverb des Satzes). Nur wenn ein Adverb sich deutlich auf den Bedeutungsbeitrag des anderen bezieht, modifizieren sie einander.

Hierbei gelten folgende Regeln:

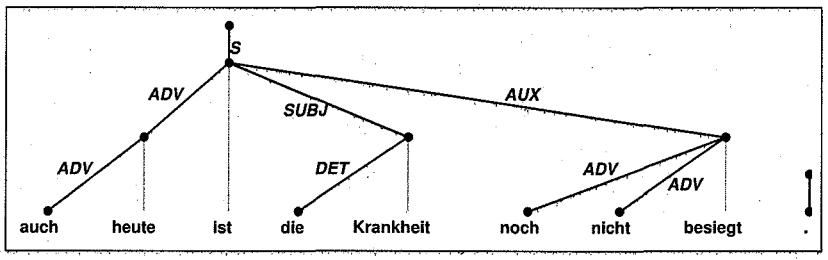
Wenn ADV zwei Adverben oder verwandte Wörter verbindet, ist meistens entweder das zweite Wort ein Adjektiv



oder das erste Wort ein Fokusadverb ('auch', 'schon', 'nur', 'erst'):

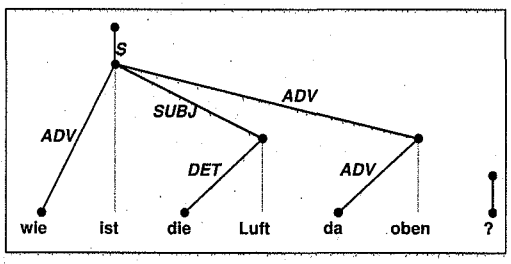


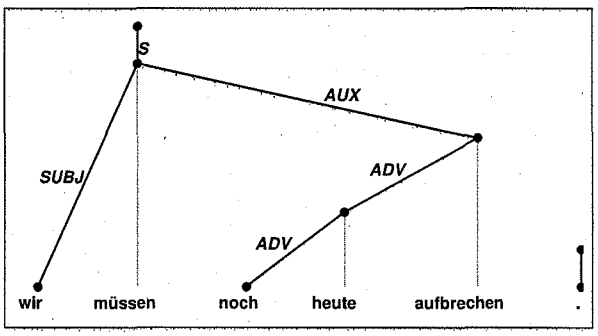
Ein Beweis dafür, daß ein Adverb das andere modifiziert, ist, daß die Adverbgruppe unverändert ins Vorfeld gestellt werden kann:



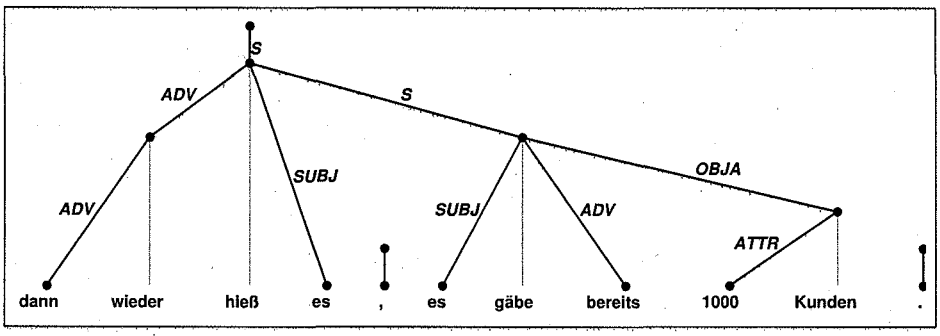
- *Nicht genug schützen sich die meisten Einwohner.

Echte Adverben modifizieren einander nur, wenn beide Zeitbedeutung oder beide Ortsbedeutung haben.

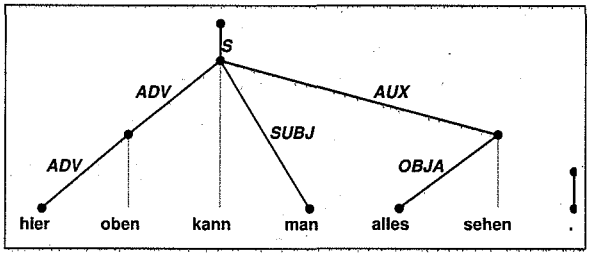




Der Vorfeldtest bestätigt:



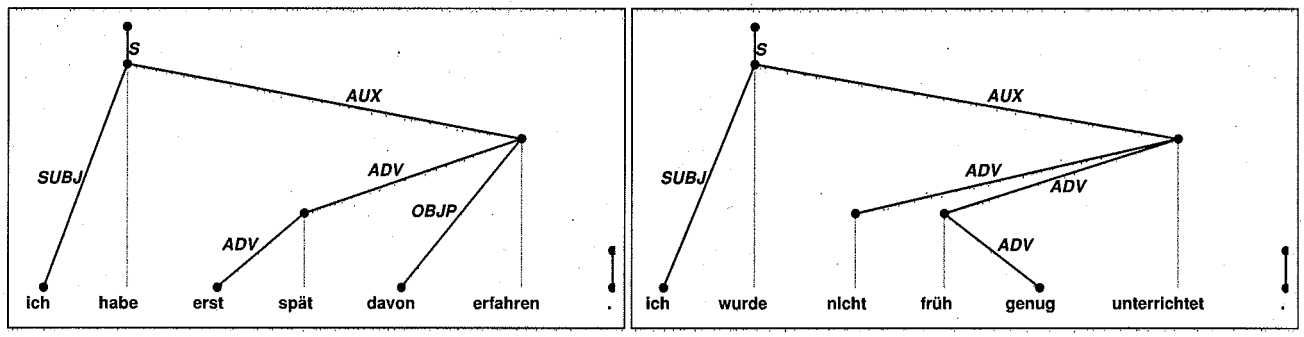
(zwei Zeitausdrücke)



(zwei Ortsausdrücke)

- *Dann oben stand er auf dem Turm. (Zeit- und Ortsausdruck)

Adverben modifizieren einander immer von links nach rechts, außer bei einigen Worten, die explizit postmodifizierend sind ('bereits', 'etwa', 'genug').

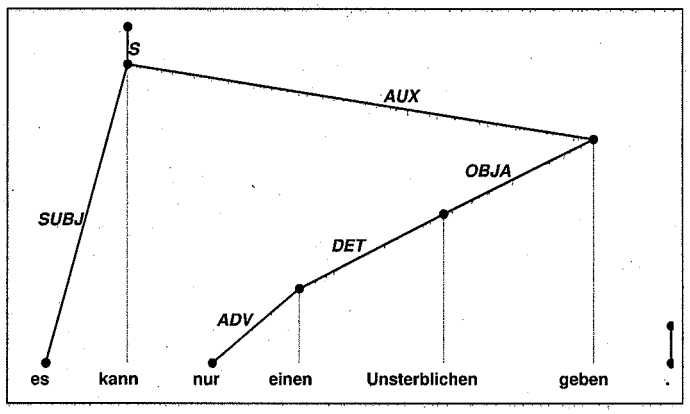


Die Fügung 'nicht mehr' wird ebenfalls auf diese Weise annotiert.

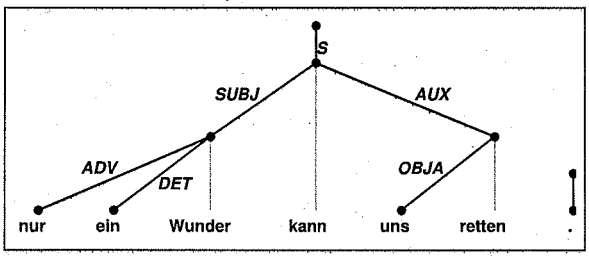
An Artikel

Im allgemeinen können nur wenige Adverben Artikel modifizieren. Typische Beispiele sind

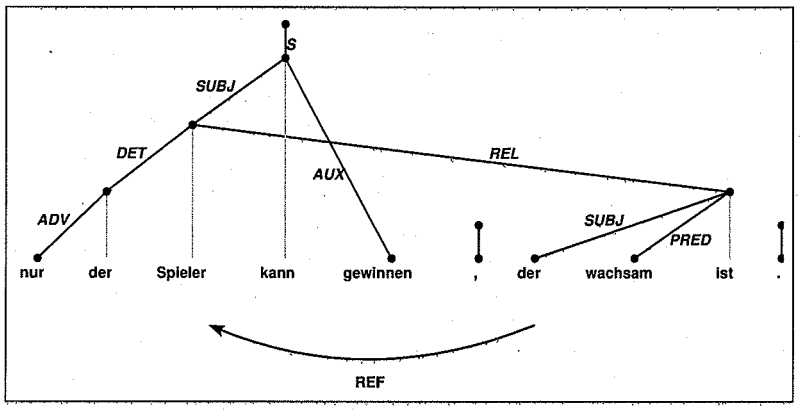
1. Fokusadverben: nur der, kaum ein... Auch hier ist aber zu unterscheiden, ob auf die gesamte NP fokussiert wird oder nur auf die Anzahl.



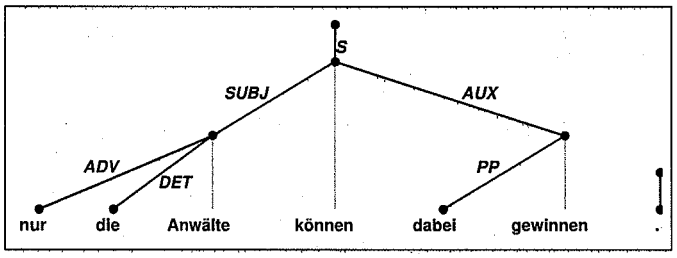
(Es kann *nicht* zwei geben, also kurze Anbindung.)



(Zwei Wunder könnten uns wahrscheinlich auch retten, also lange Anbindung.)

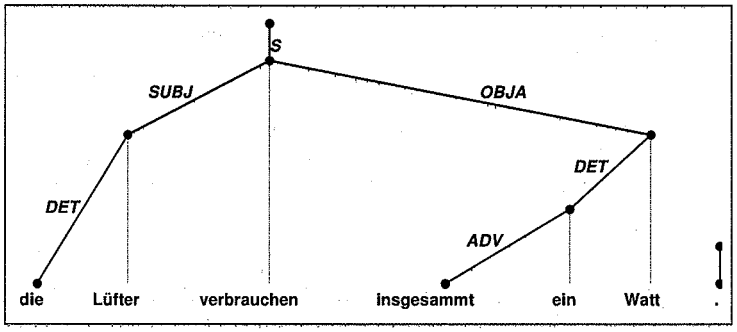


(Ein anderer Spieler kann *nicht* gewinnen, also kurz.)

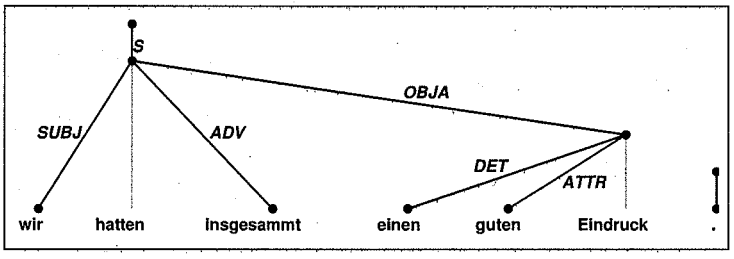


(Andere Anwälte können auch gewinnen, nur nicht die Betroffenen, also lang.)

2. Adverben an 'ein' mit Zahlbedeutung: mindestens ein, etwa ein.... Auch hier muß geprüft werden, ob nicht die ganze VP modifiziert wird:

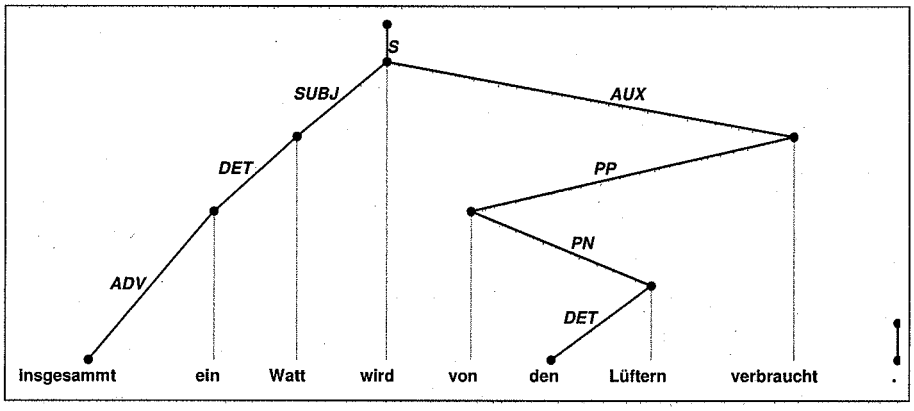


(Zwei Watt wären auch denkbar, also kurze Anbindung.)



(Zwei gute Eindrücke sind nicht möglich, also lange Anbindung.)

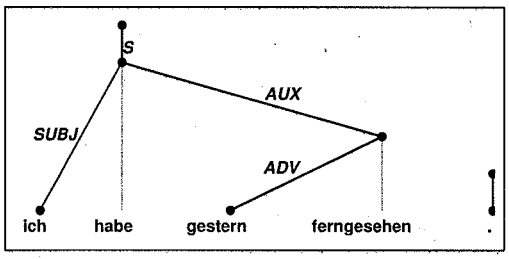
Hier kann auch der Vorfeldtest angewendet werden: wenn das Adverb den Artikel modifiziert, kann die gesamte Konstruktion vor das Verb gesetzt werden, sonst nicht.



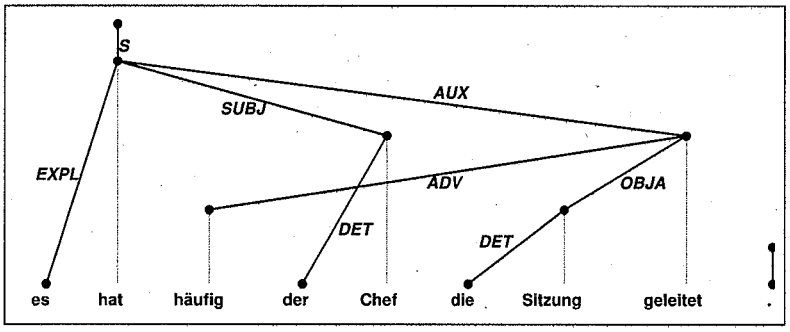
- *Insgesamt einen guten Eindruck hatten wir.

An Vollverb oder Hilfsverb?

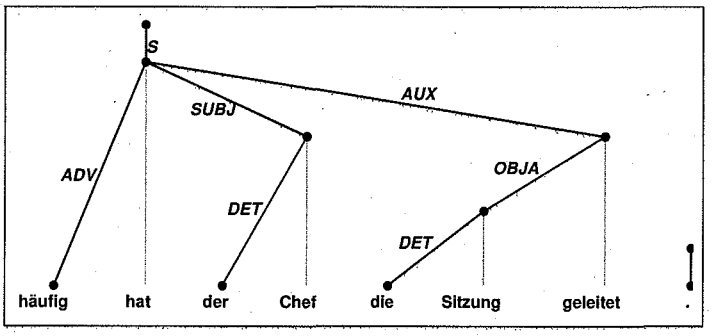
Tritt ein Adverb zwischen den Bestandteilen der Verbklammer auf, also im Mittelfeld, soll es stets an das Vollverb angebunden werden.



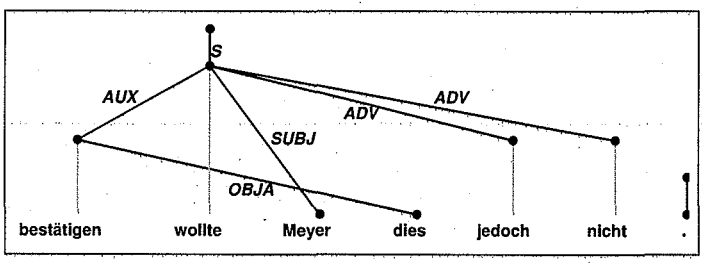
Das gilt selbst dann, wenn es dabei das Subjekt überkreuzt:



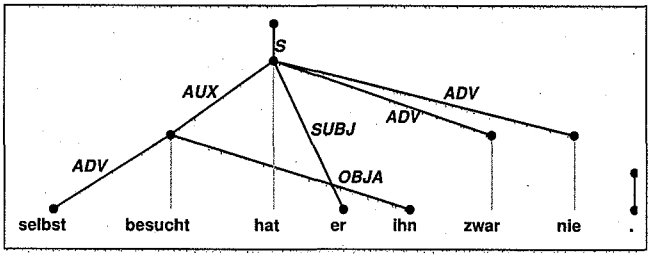
In allen anderen Fällen wird die kürzeste mögliche Anbindung gewählt. Steht das Adverb im Vorfeld, so modifiziert es das finite Verb:



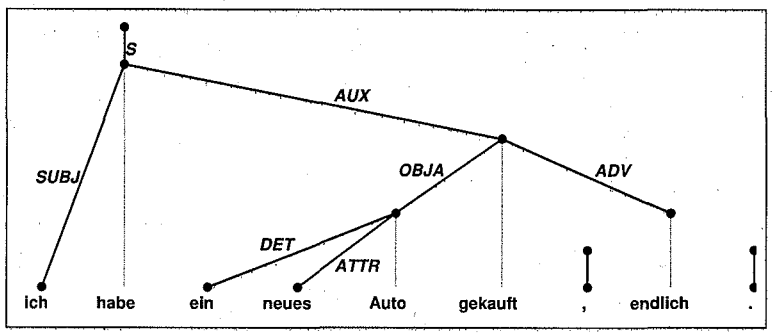
Steht das Vollverb im Vorfeld und das Adverb im Nachfeld, so modifiziert das Adverb das finite Verb:



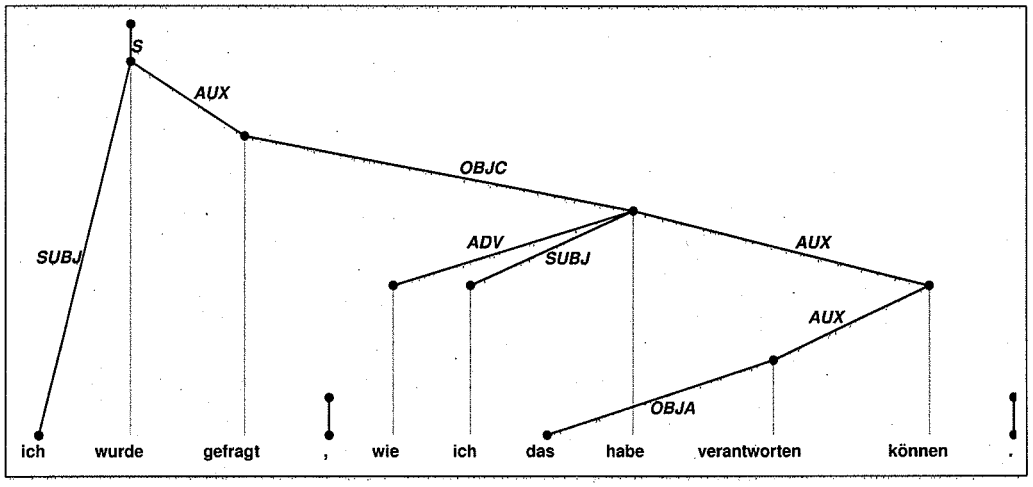
Stehen aber sowohl Adverb als auch Vollverb im Vorfeld, so modifiziert das Adverb das Vollverb:



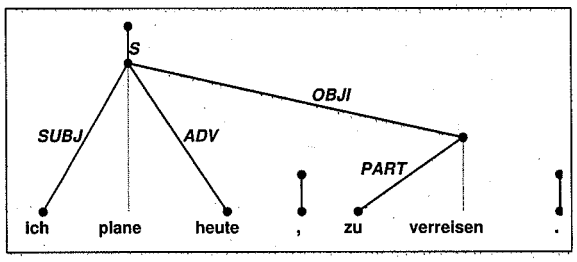
Steht das Adverb im Nachfeld, so modifiziert es das Vollverb:



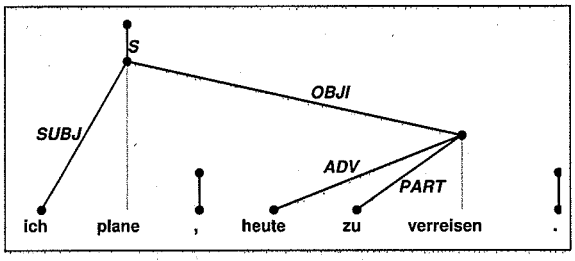
Liegt eine Inversion in der Verbgruppe vor, so wird das finite Verb modifiziert:



Verbgruppen, die nicht mit Hilfsverben, sondern durch Vollverben gebildet werden, erlauben die Unterordnung sowohl links als auch rechts, je nach der Satzbedeutung.



(Die Reise findet morgen statt.)



(Die Reise findet heute statt.)

Richtlinie: wenn ein Komma den Nebensatz markiert, muß das Adverb diesseits des Kommas angebunden sein.

An Nomen oder Verb?

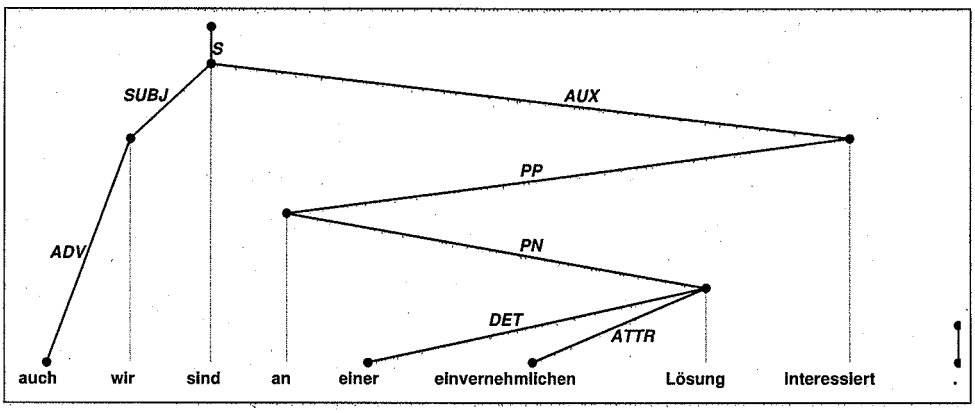
Verschiedene Adverben modifizieren sowohl NP als auch VP und adverbiale Bestimmungen. Insbesondere bei den fokussierenden Adverben wie 'auch' oder 'nicht' ergeben sich dabei Sinnänderungen, die nicht unberücksichtigt bleiben sollten:

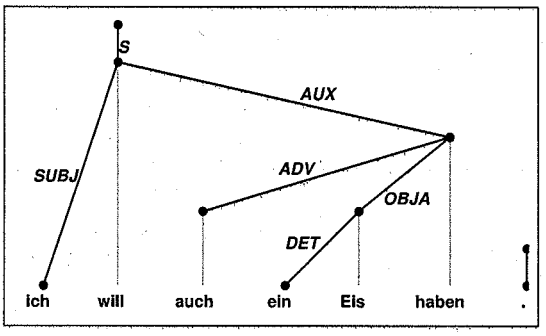
- Hatten Sie auch mit Feindagenten Kontakte?

Je nach der Satzbedeutung sollte hier die passendste Unterordnung gewählt werden:

Präsupposition	Regent von 'auch'
Sie arbeiteten im Strafvollzug.	'hatten'
Sie hatten Kontakte.	'mit'
Ihr Mann hatte solche Kontakte.	'Sie'

Nicht in allen Fällen kann diese Regel eingehalten werden. Beispielsweise kann ein fokussiertes Element im Vorfeld *nicht* aus dem Mittelfeld modifiziert werden. Nur das erste der beiden nächsten Beispiele kann also sinngemäß annotiert werden:



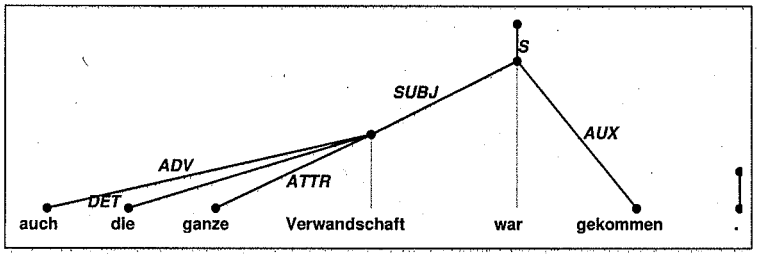


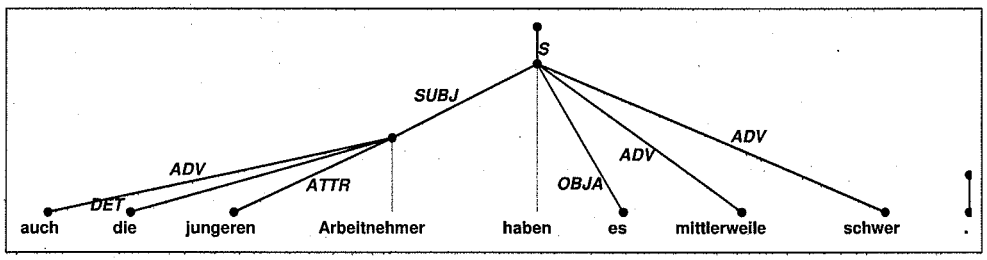
Im zweiten Fall kann das 'auch' nicht das inhaltlich richtige 'ich' modifizieren.

An Adjektiv oder Nomen?

Steht ein Adverb vor einer NP, die noch attributive Adjektive enthält, so könnte es sowohl das Nomen ('lang') als auch das Adjektiv ('kurz') modifizieren. Es sind bei der Entscheidung verschiedene Kriterien zu beachten.

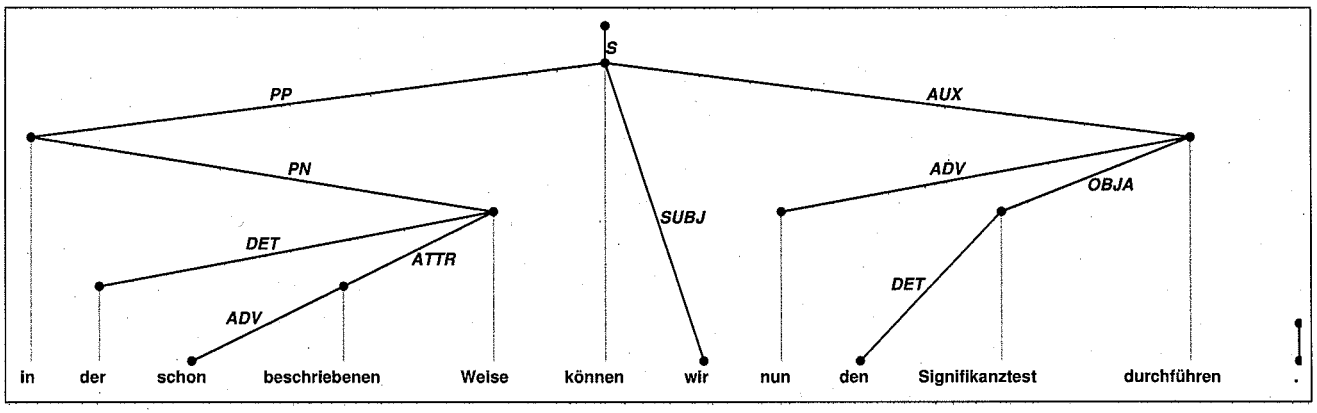
Zunächst gibt es rein strukturelle Gründe. Wenn zwischen Adverb und Nomen ein Determiner auftritt, so modifiziert das Adverb immer das Nomen:



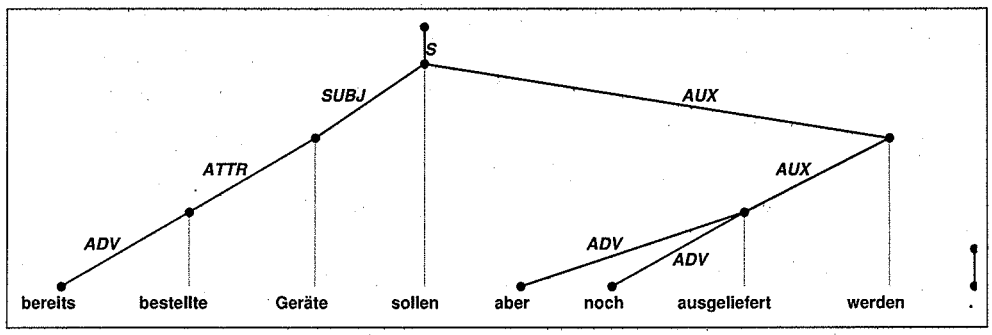


(Das gilt auch dann, wenn es inhaltlich enger zum Nomen gehört wie im zweiten Beispiel oben: der Kontrast ist 'ältere Arbeitnehmer' und nicht 'Arbeitgeber'.)

Tritt umgekehrt das Adverb zwischen Determiner und Adjektiv, so modifiziert das Adverb immer das Adjektiv:

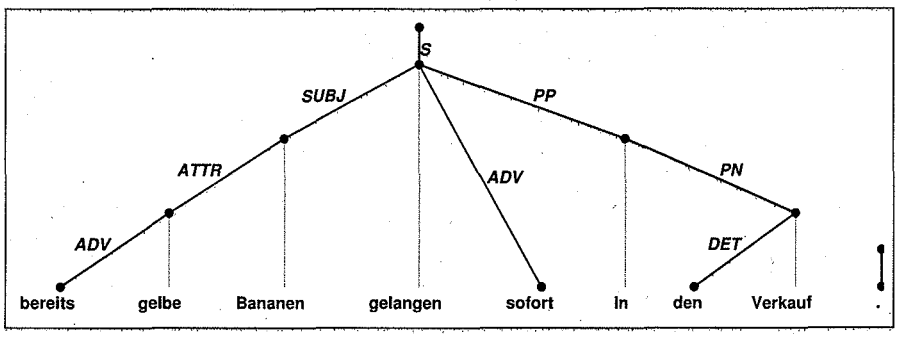


Wenn kein Determiner vorhanden ist, sind beide Alternativen strukturell möglich. In diesem Fall muß anhand der Satzbedeutung entschieden werden. Wenn der durch das Adjektiv implizierte Basis-Satz ebenfalls das Adverb enthält, so ist das Adverb an das Adjektiv anzubinden, anderenfalls an as Nomen. Beispiel:

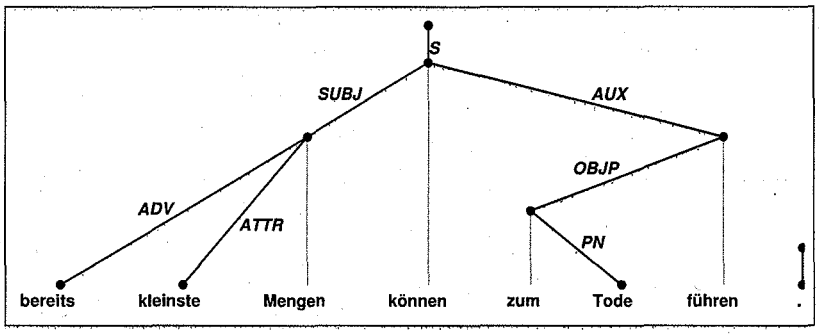


(Präsupposition: Es wurden bereits Geräte bestellt, also kurze Anbindung.)

Der Satz impliziert 'Es wurden bereits Geräte bestellt', daher modifiziert 'bereits' das Adjektiv 'bestellte'. Diese Art der Konstruktion ist nicht auf Partizipialadjektive beschränkt:



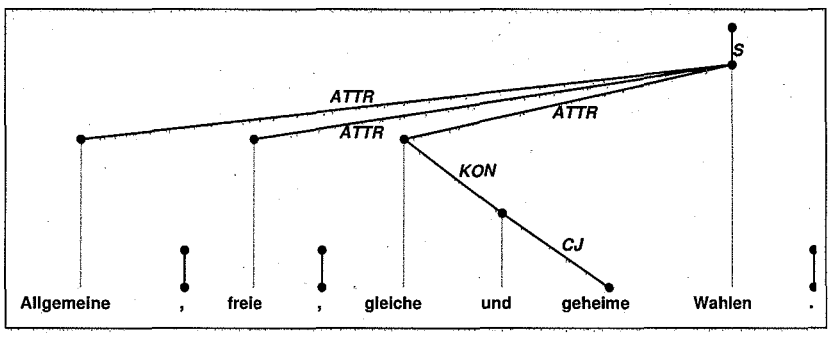
(Präsupposition: Die Bananen sind bereits gelb.)



(Die Präsupposition ist *nicht* 'Die Mengen sind bereits kleinst'.)

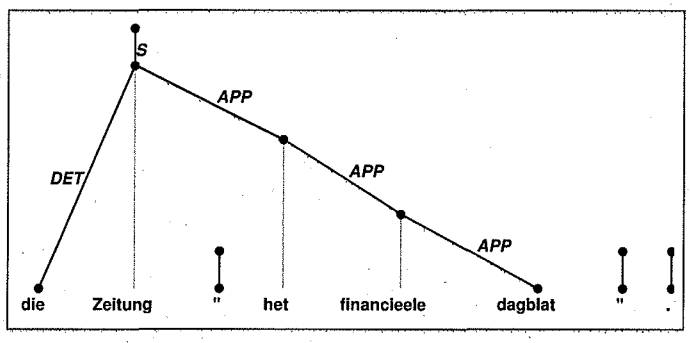
Anbindung von ATTR

Stehen mehr als zwei Adjektive am selben Nomen, so sind sie alle nebenzuordnen. Wenn eine Konjunktion auftritt, verbindet sie nur die beiden benachbarten Adjektive, die anderen modifizieren direkt das Nomen.

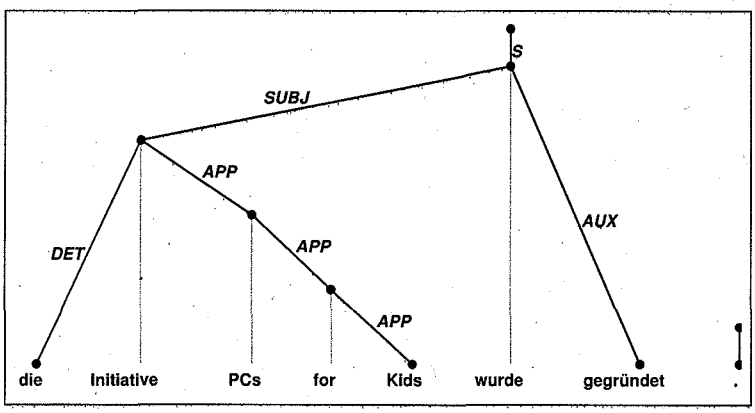


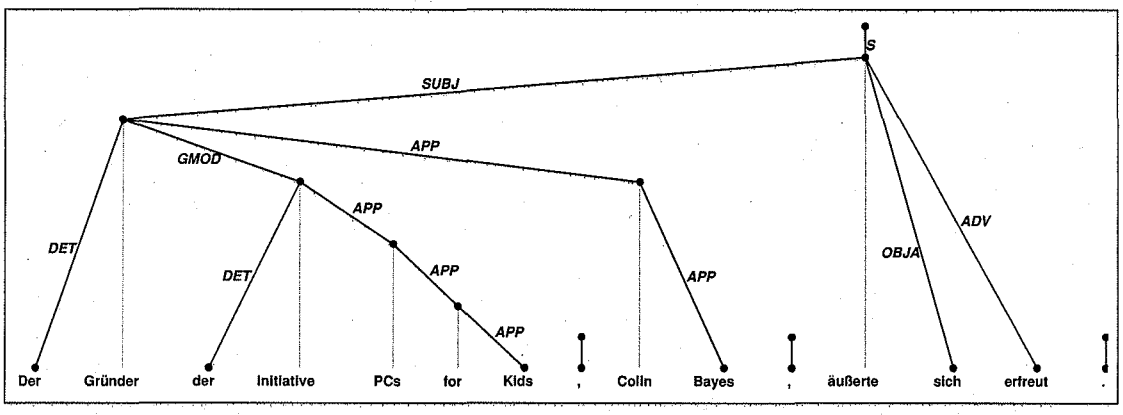
Anbindung von APP

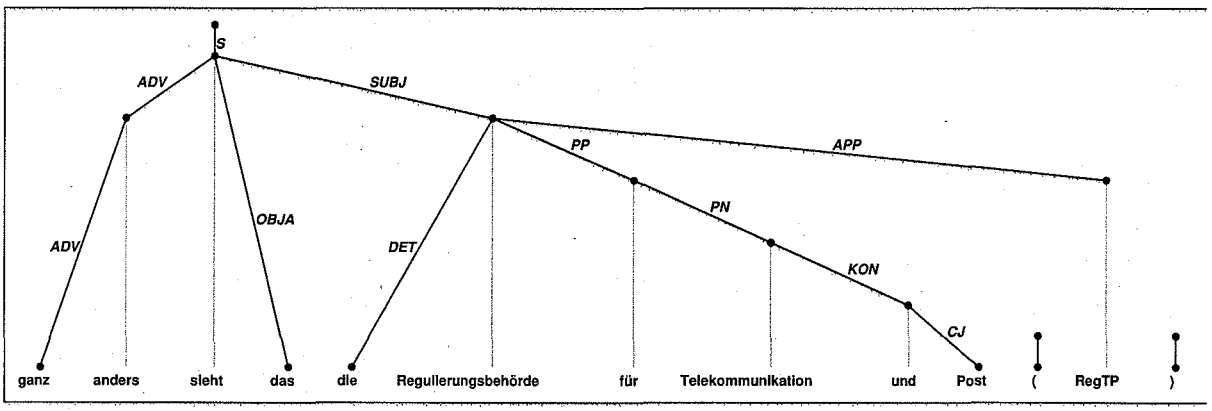
Wenn mehrere Worte eine NP bilden, modifiziert jedes Hauptwort (NN, NE, FM) das vorige Hauptwort. Das gilt auch, wenn es sich um fremdsprachliche Worte handelt, die erkennbar eine andere Struktur haben.



NP, die nur durch APP gebildet werden, verzweigen also niemals. Anders sieht es aus, wenn eine komplexe NP Genitivattribute, Präpositionen oder gar Nebensätze enthält. Soweit erkennbar von verschiedenen Personen oder Dingen die Rede ist, muß hier jede Teil-NP diejenige modifizieren, deren Wiederaufnahme sie ist.







Hier muß der Syntaxbaum also ausdrücken, daß Colin Bayes der Gründer ist (nicht das Kid), und daß RegTP für 'Regulierungsbehörde' steht und nicht für 'Post'.

Anbindung von AUX

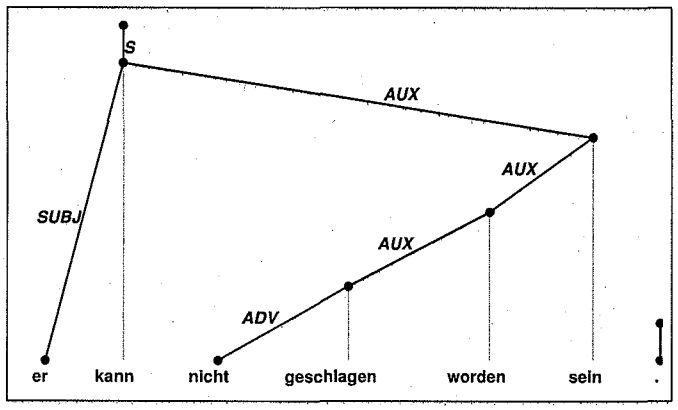
Die Reihenfolge der Unterordnung in der Auxiliargruppe richtet sich danach, welche Funktion die einzelnen Worte haben:

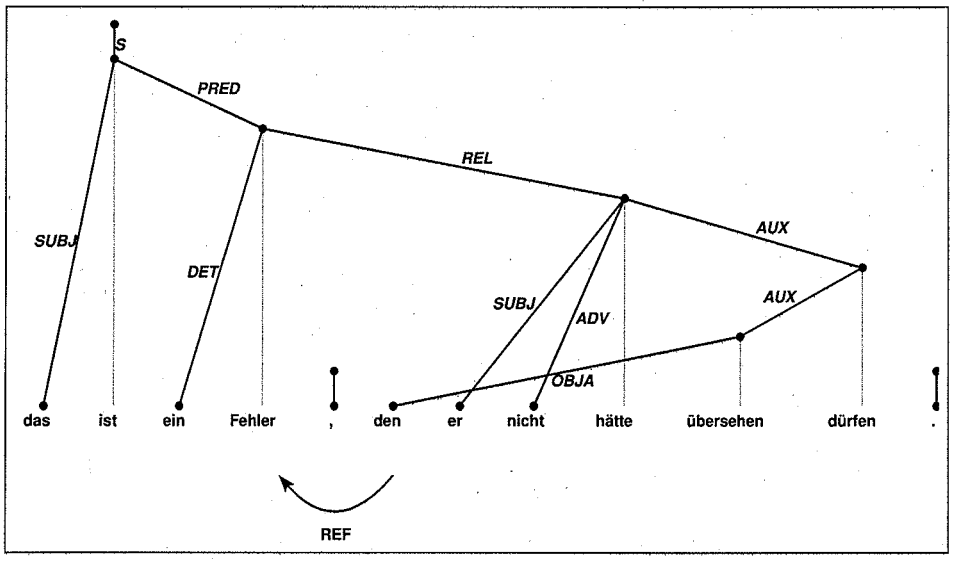
1. Tempus (VAFIN, VMFIN)
2. Modalität (VMINF)
3. Passiv (VAINF, VAPP von 'sein', 'werden')
4. Inhalt (VVINF, VVPP)

Die normale Reihenfolge ist also

- Als er geschlagen→ worden→ war...

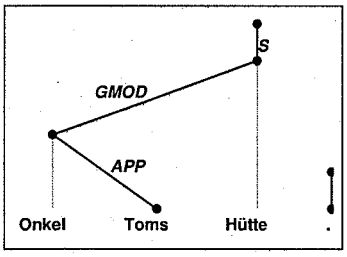
Diese Struktur wird auch dann beibehalten, wenn die Verbgruppe umgestellt wird, selbst dann, wenn der Syntaxbaum dadurch nichtprojektiv wird.





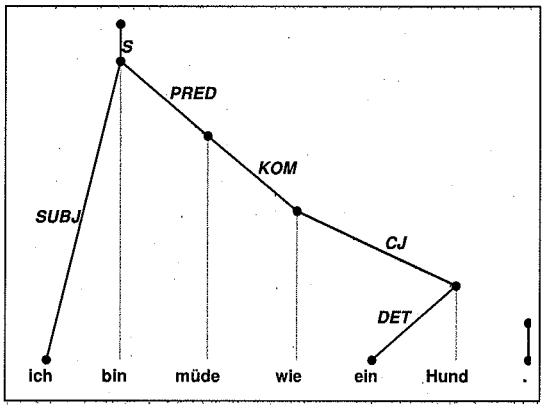
Anbindung von GMOD

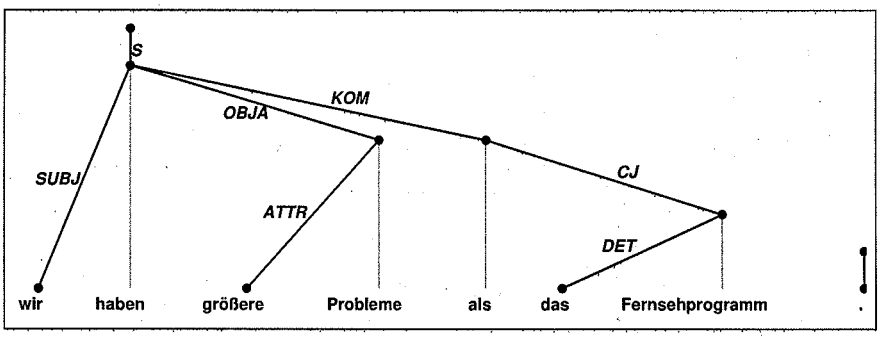
Wenn ein mehrteiliger Eigenname als Genitivmodifikator verwendet wird, ist dennoch der linke Bestandteil GMOD und die anderen APP. In diesem Fall kann also auch ein Wort GMOD sein, das selbst eindeutig im Nominativ steht.

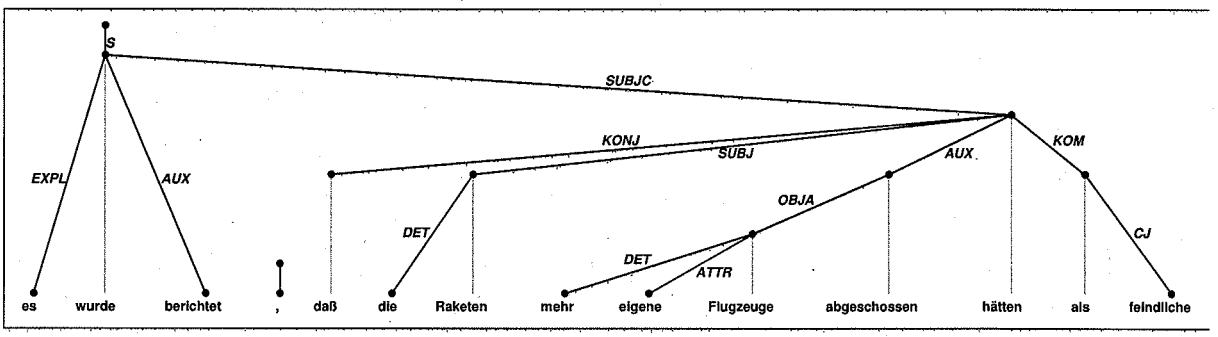


Anbindung von KOM

Vergleichsaussagen werden häufig durch einen Positiv (mit 'wie') oder Komparativ (mit 'als') ausgelöst, der tief in den Hauptsatz verschachtelt ist. Für diese Erscheinung wird keine Ausnahme von der Projektivität gemacht. Die KOM-Modifikation muß also so weit wie nötig angehoben werden.

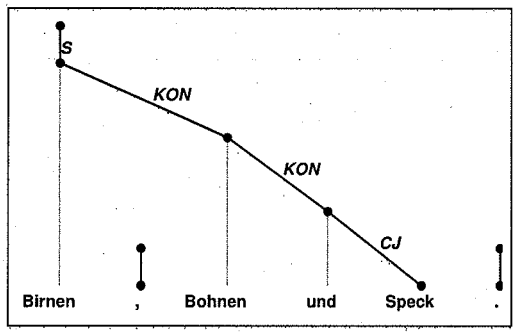


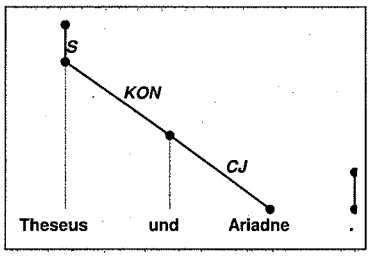




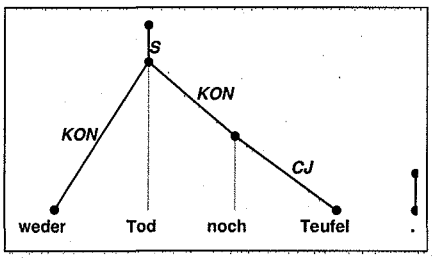
Anbindung von KON

In Nebenordnungen ist immer das erste Wort dem zweiten übergeordnet, egal ob eine Konjunktion auftritt oder nicht.

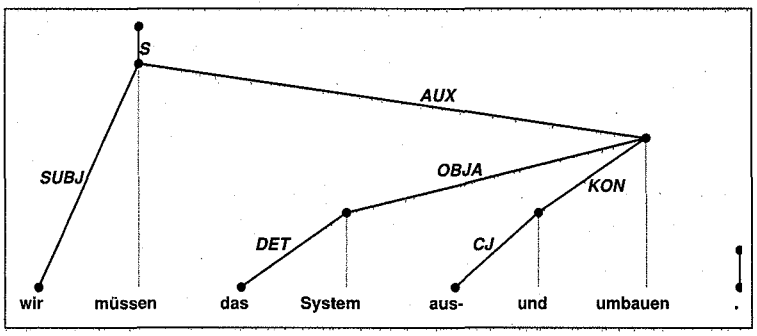




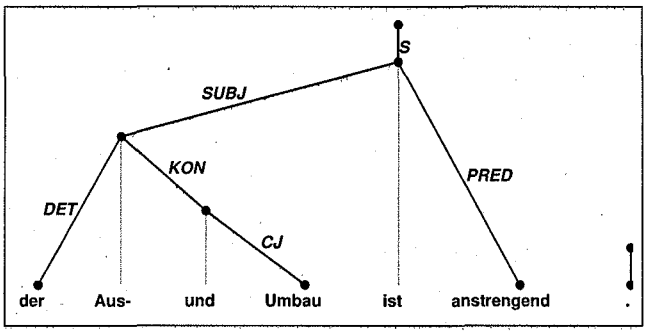
Es gibt nur zwei Situationen, in denen die Beziehung andersherum verläuft. Zum einen sind das die linken Teile von zusammengesetzten Konjunktion ('weder', 'entweder'):

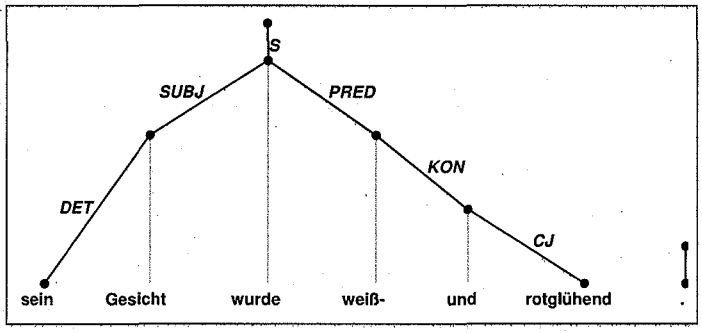


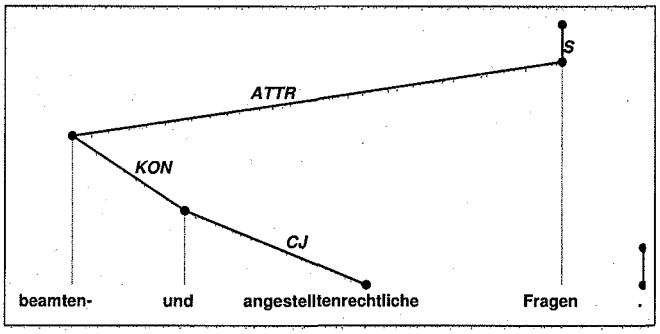
Zum anderen werden verkürzte Verben dem vollständigen Verb untergeordnet, obwohl sie links stehen:



Dies gilt nur für Verben. Andere verkürzte Wortteile werden normal behandelt:



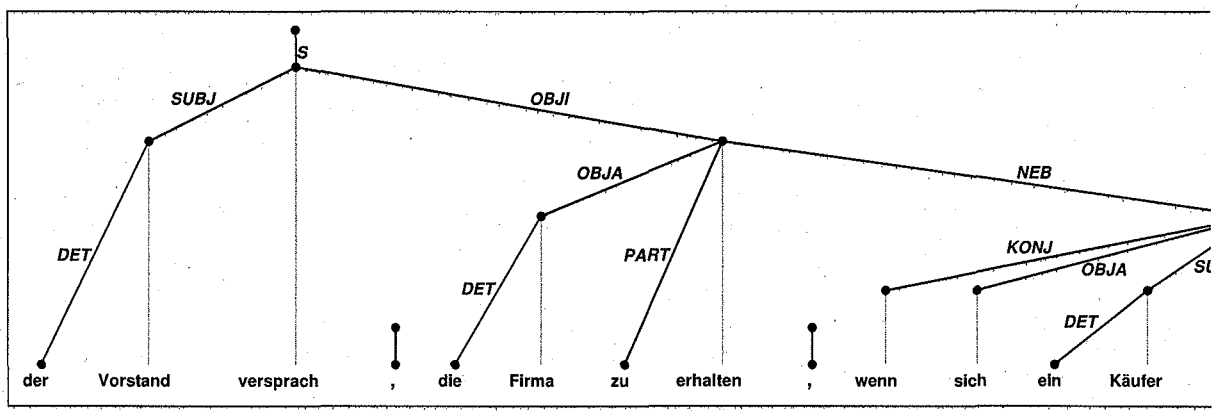




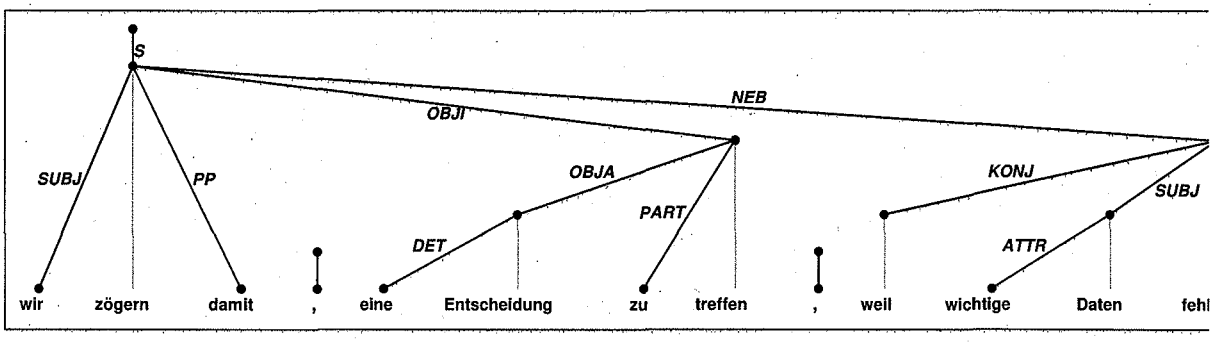
Anbindung von NEB

Nebensätze, die von komplexen Verbgruppen abhängen, werden hoch angebunden, also an das finite Verb. Es wird nur dann das Vollverb modifiziert, wenn das finite Verb nicht vorhanden oder z.B. hinter einer Koordination unzugänglich ist.

Sind mehrere finite Verben erreichbar, so entscheidet die Satzbedeutung, ob die kurze oder lange Anbindung zu wählen ist.

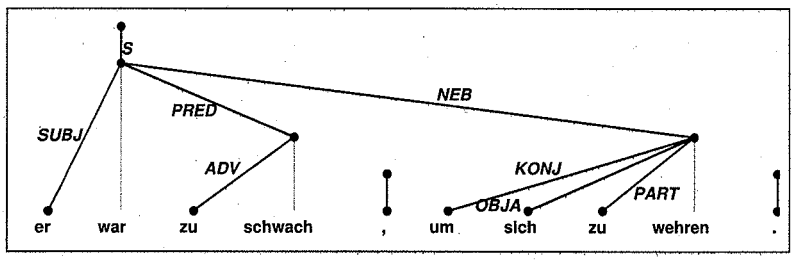


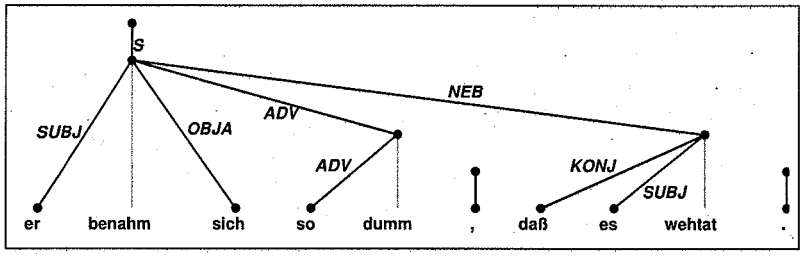
(Der Erhalt hängt vom Käufer ab, das Versprechen nicht, also kurze Anbindung.)

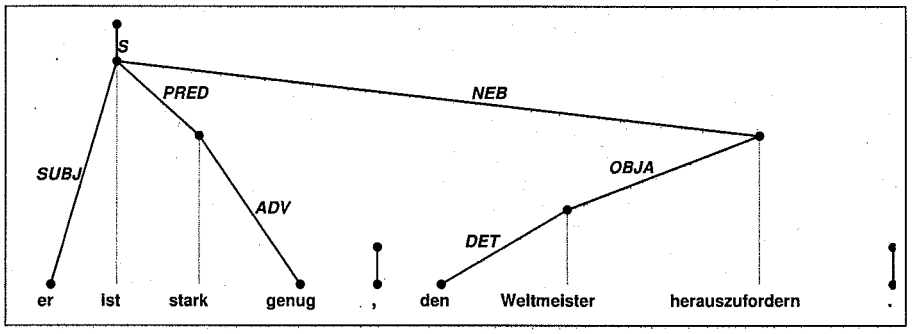


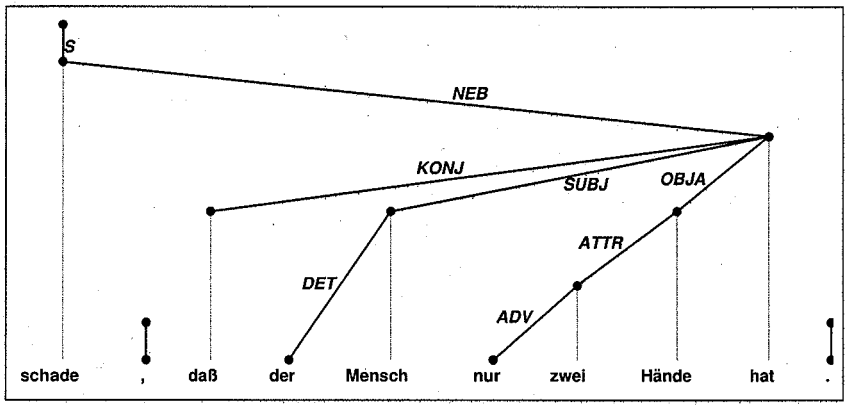
(Das Fehlen begründet das Zögern, nicht die Entscheidung, also lange Anbindung.)

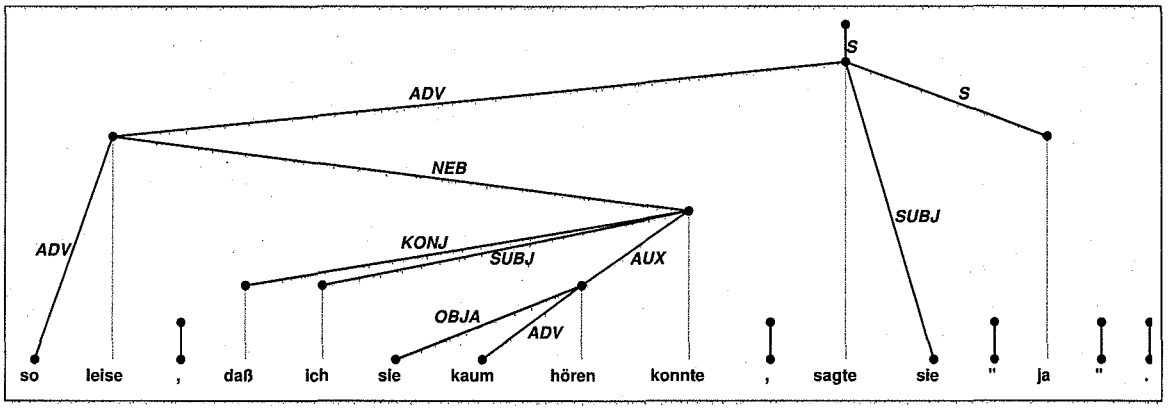
Dasselbe gilt von der Anbindung an Adjektive, die durch 'zu', 'genug' oder durch intentionale Bedeutung einen Nebensatz lizensieren: der Nebensatz modifiziert das Hauptsatzverb, wenn dies erreichbar ist, und das Adjektiv nur dann, wenn das Hauptsatzverb fehlt oder hinter einer Koordination 'verborgen' ist. Also:

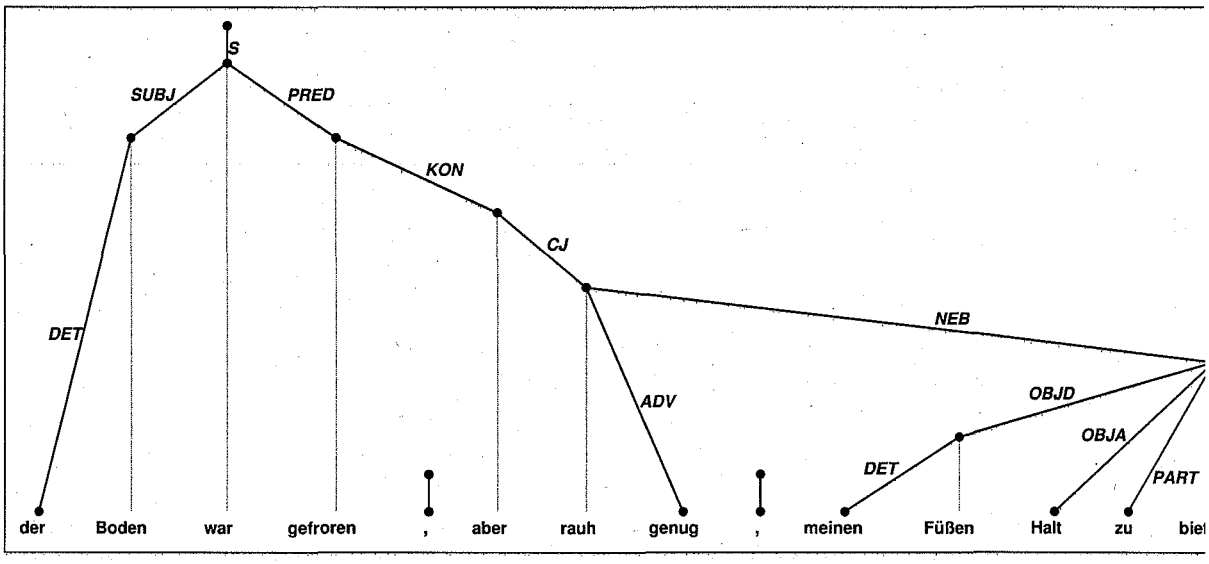

S
NEB
PRED
SUBJ
ADV
KONJ
OBJA
PART
er war zu schwach , um sich zu wehren .


S
NEB
SUBJ
OBJA
ADV
ADV
KONJ
SUBJ
er benahm sich so dumm , daß es wehtat .


S
PRED
NEB
SUBJ
OBJA
ADV
DET
er ist stark genug , den Weltmeister herauszufordern .

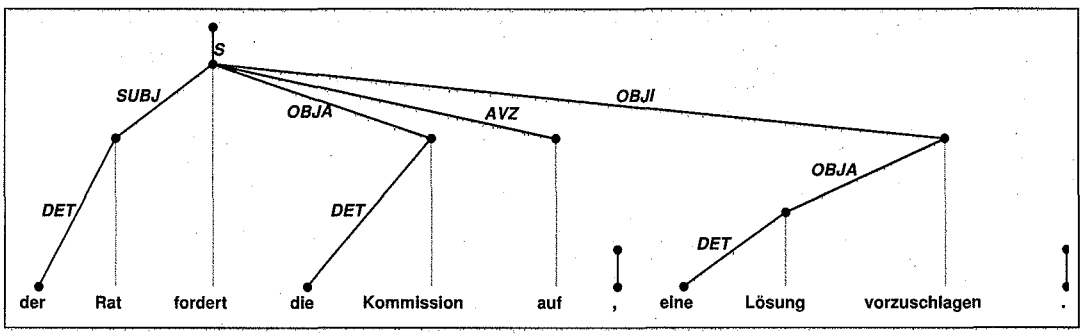

S
NEB
KONJ
SUBJ
OBJA
ATTR
DET
ADV
schade , daß der Mensch nur zwei Hände hat .



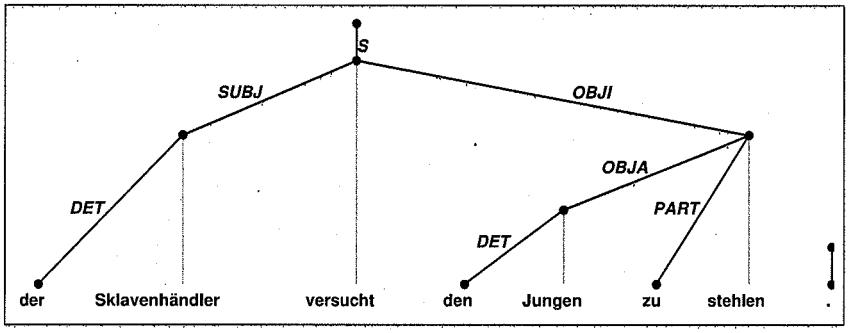


Anbindung von OBJA

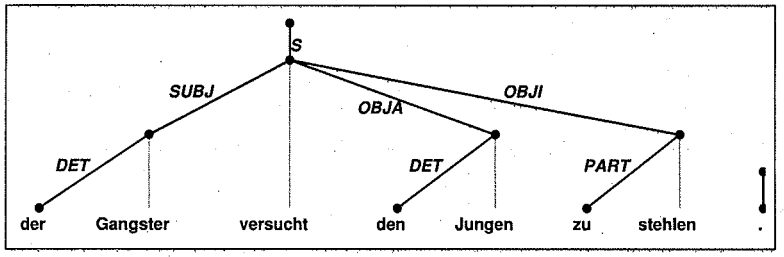
Das Akkusativobjekt modifiziert stets dasjenige Vollverb, zu dem es inhaltlich gehört.



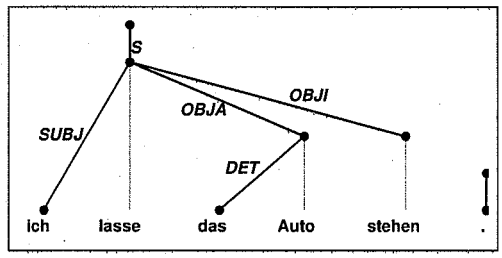
Bei bestimmten Verben sind beise Varianten möglich, jedoch mit Sinnänderung; in diesem Fall muß die wahrscheinlich bebabsichtigte Bedeutung berücksichtigt werden. Das Subjekt des untergeordneten Satzes modifiziert das übergeordnete Verb, das Objekt des modifiziert das andere Vollverb:



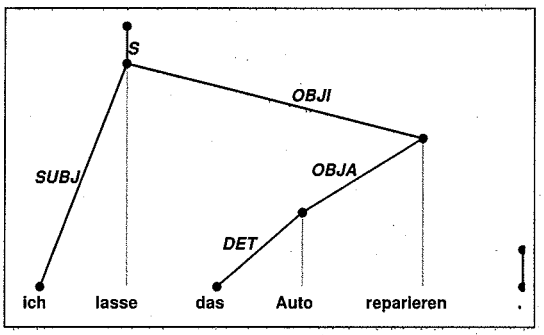
(Der Sklavenhändler will stehlen.)



(Der Junge soll stehlen.)

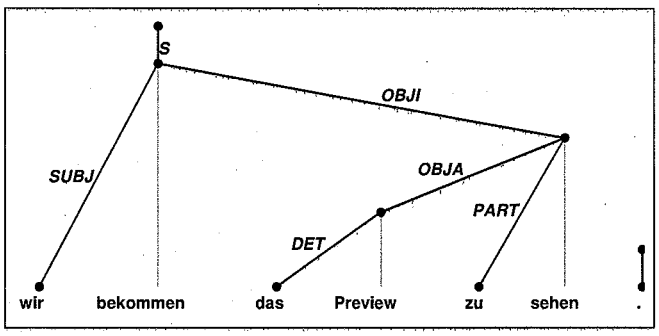


(Das Auto steht.)



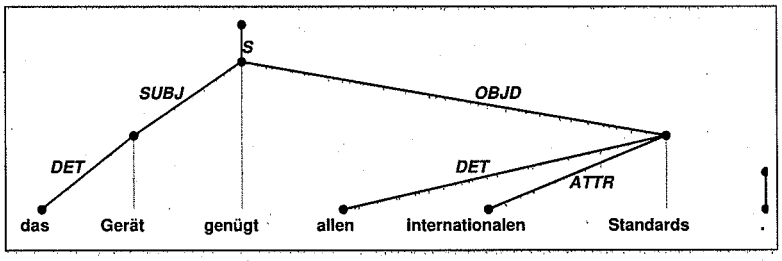
(Jemand anders repariert das Auto).

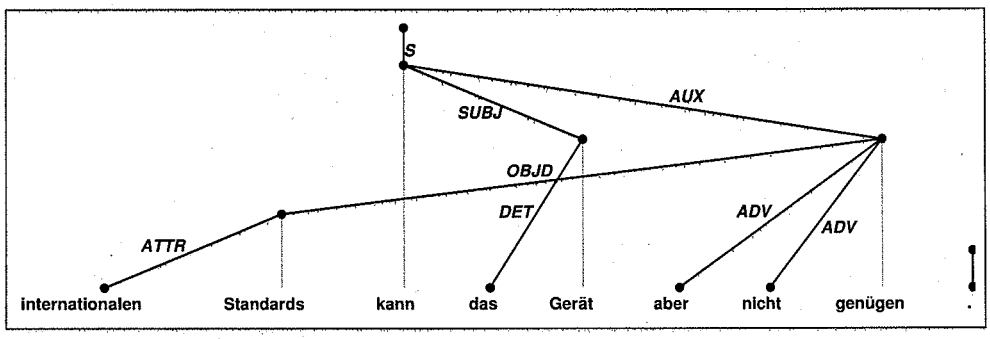
Wenn beide Möglichkeiten richtig erscheinen, soll die tiefe Anbindung gewählt werden:



Anbindung von OBJD

Wenn das Dativobjekt eine Valenz des Verbs ist, modifiziert es stets das Vollverb. Ist es dagegen ein ethischer (spontaner) Dativ, so gelten dieselben Regeln wie für adverbiale Bestimmungen, d.h. es wird stets die kurze Anbindung gewählt.

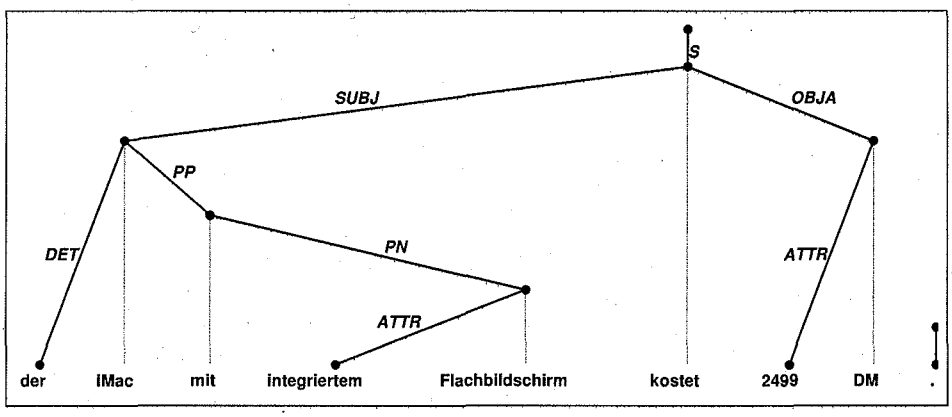


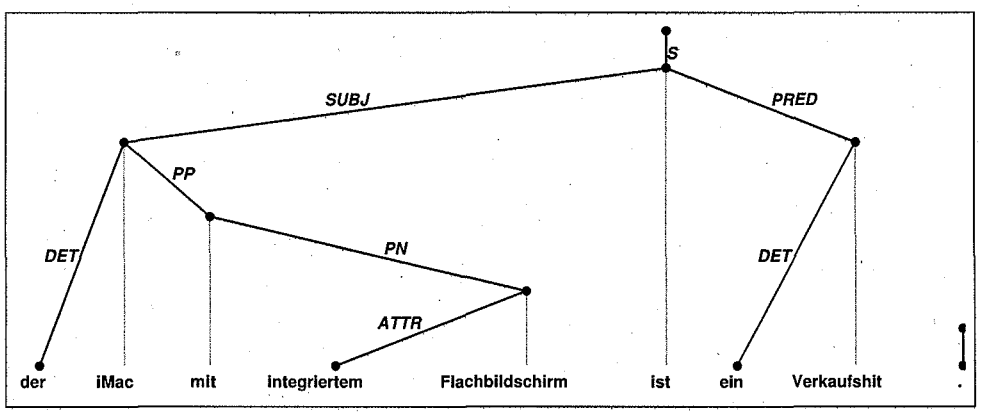


Anbindung von PP

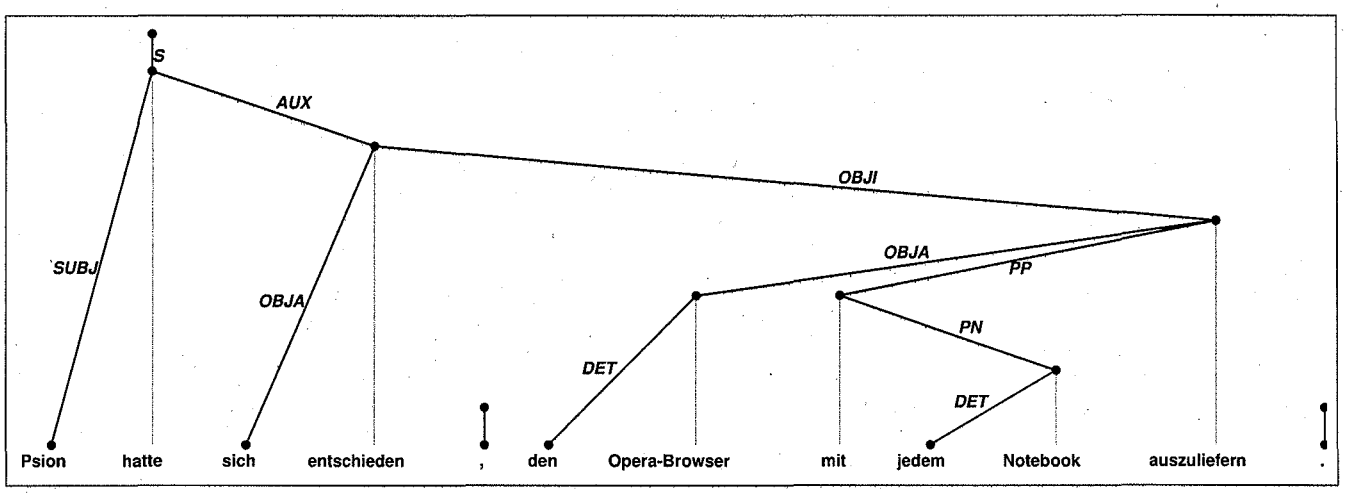
An Verb oder Nomen?

Wenn sowohl NP als auch VP als Anbindungspunkt in Frage kommen, so muß die inhaltlich sinnvolle Variante gewählt werden. Insbesondere sollte die Anbindung an eine NP nur dann gewählt werden, wenn die NP + PP auch in anderem Zusammenhang sinnvoll wäre:





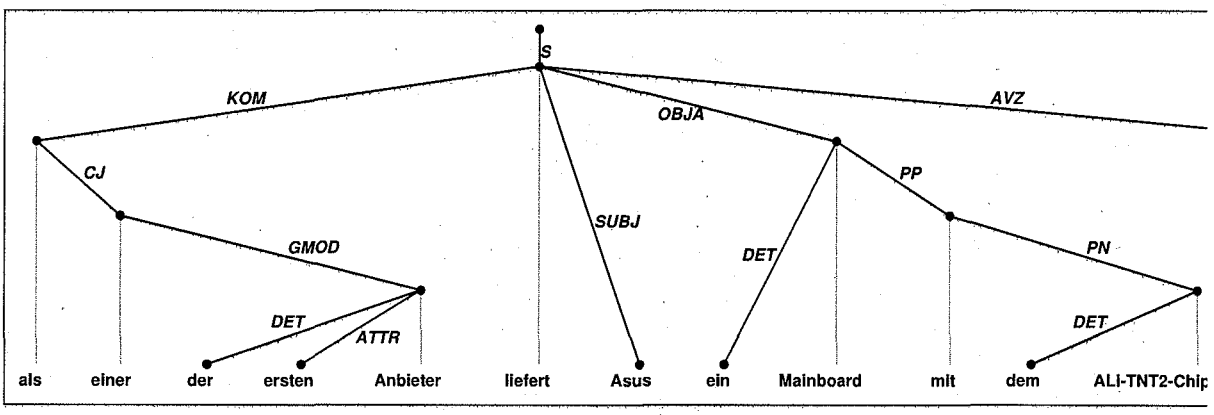
(Es ist charakteristisch für den iMac, daß sein Bildschirm integriert ist.)



- *Der Opera-Browser mit jedem Notebook hat die Version 4.53 erreicht.

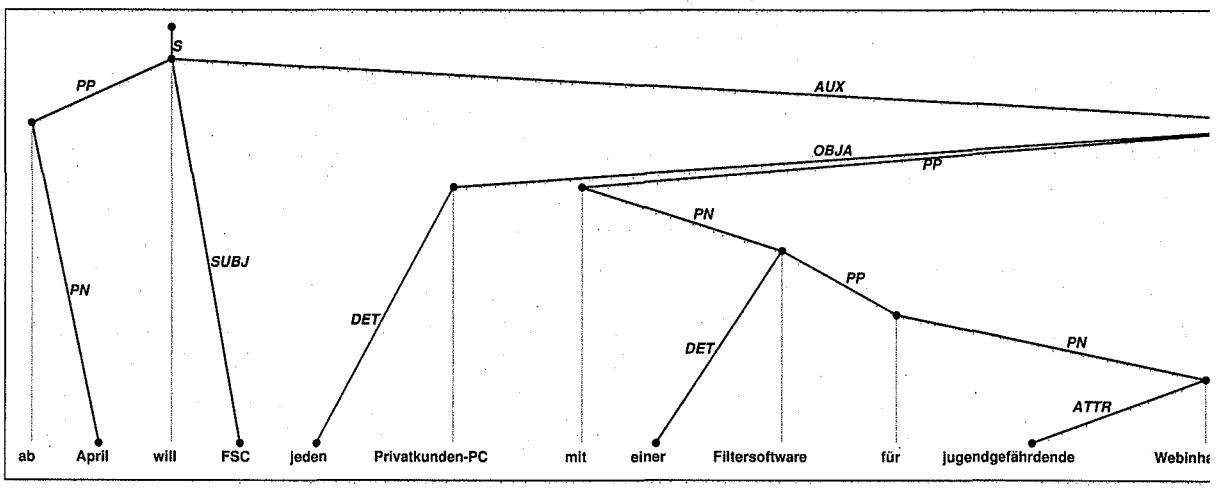
(Es ist nicht charakteristisch für den Browser, daß er ein Notebook enthält.)

Diese Entscheidung setzt natürlich Wissen über den Satzgegenstand voraus. Beispielsweise ist im folgenden Beispiel die Nomenanbindung zu wählen:



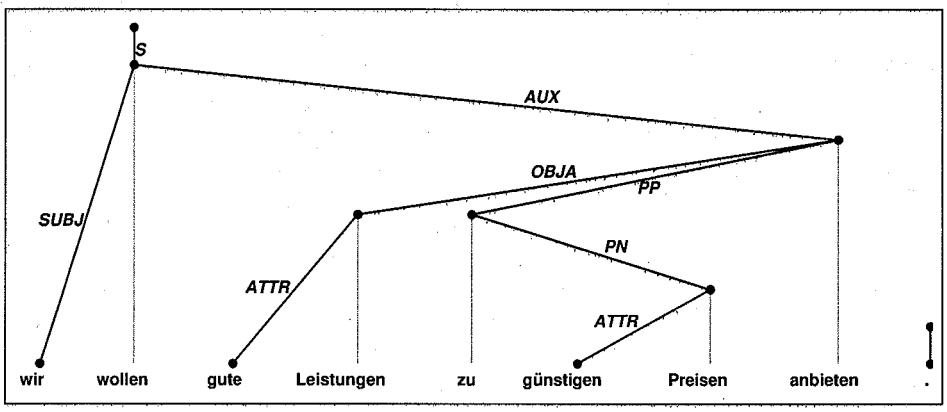
Das ist deshalb der Fall, weil der bewußte Chipsatz der wesentliche, fest eingebaute Hauptbestandteil des Produktes ist. Die Anbindung and das Verb 'liefert' würde ausdrücken, daß Mainboard und Chipsatz als getrennte Produkte sind, aber im selben Karton versandt werden. Es kann notwendig sein, den gesamten Text oder sogar Hintergrundinformationen Material zu lesen, um eine solche Entscheidung zu treffen. Wenn solche satzexterne Information dazu beiträgt, die Entscheidung zu treffen, sollte sie auch angewendet werden (obwohl der Parser sie natürlich nicht nutzen kann).

Im folgenden Beispiel dagegen ist die Verbanbindung zu wählen:

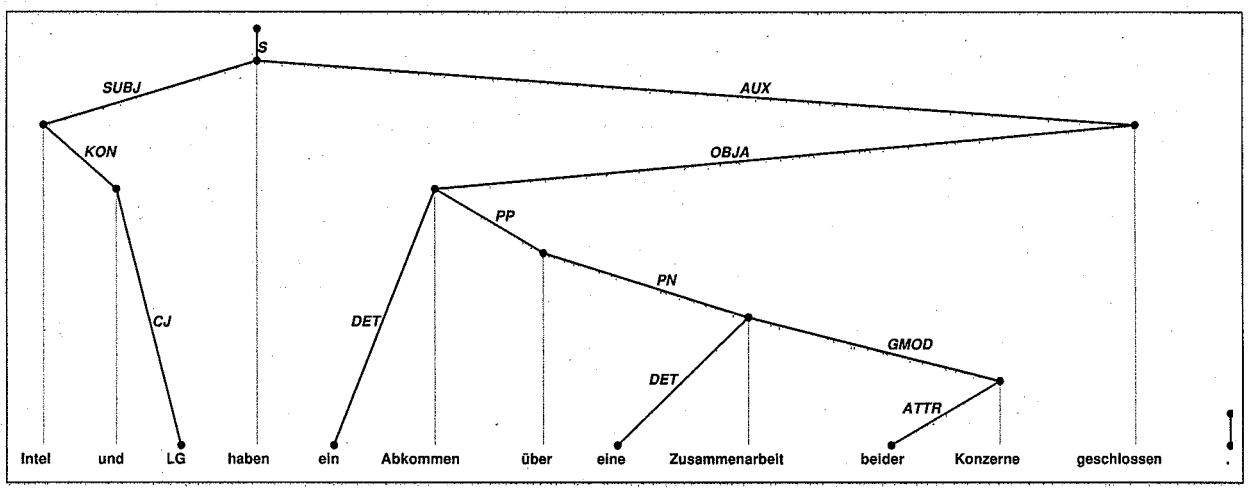


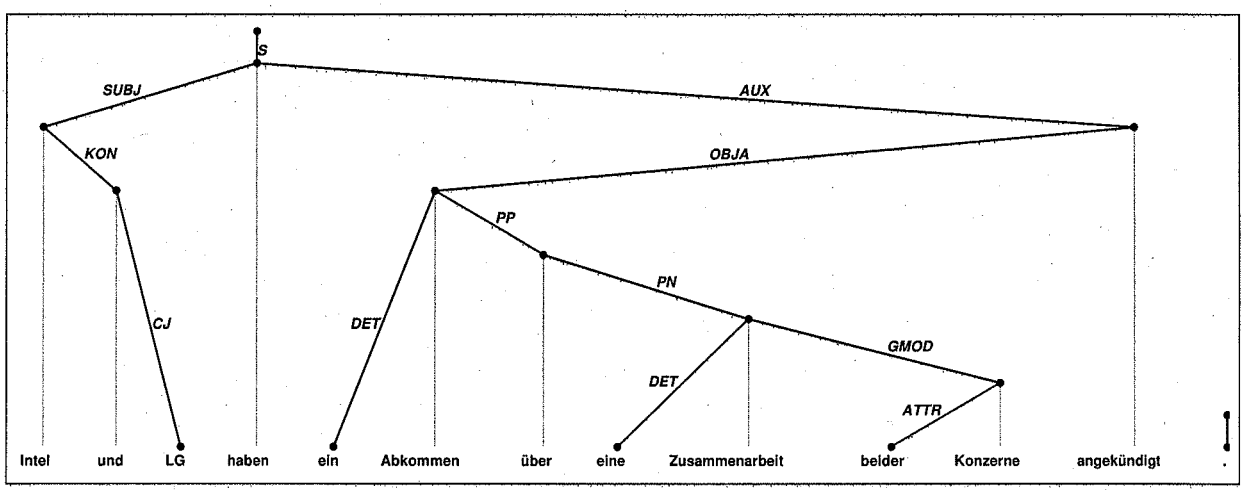
Aufgrund des Quantors 'jeden' würde die Nomenanbindung bedeuten, daß die Firma weiterhin PCs mit und ohne Filtersoftware herstellt, aber diejenigen ohne Filtersoftware nicht mehr ausliefert. Gemeint ist natürlich, daß künftig jedem PC eine Filtersoftware beigefügt wird.

Wenn kein Bedeutungsunterschied erkannt werden kann, soll die Anbindung ans Verb gewählt werden:



Ein guter Test ist oft die Ersetzbarkeit der beteiligten Nomen und Verben:





- *Intel und LG haben eine Fabrik über eine Zusammenarbeit beider Konzerne geschlossen.

Hier läßt sich durch gezielte Ersetzung des Nomens die Akzeptabilität des Satzes deutlich

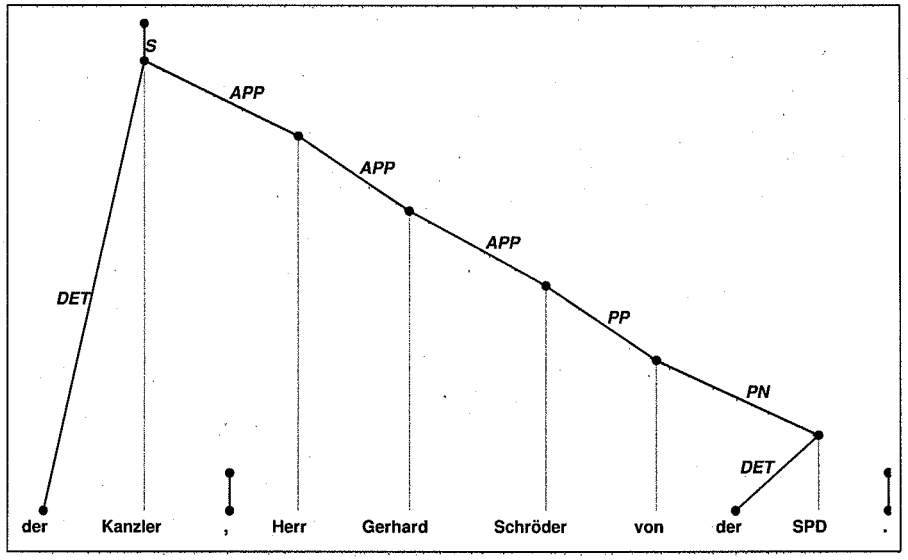
verringern, während die Ersetzung des Verbs keine Auswirkung hat. Das deutet drauf hin, daß die PP das Nomen modifiziert.

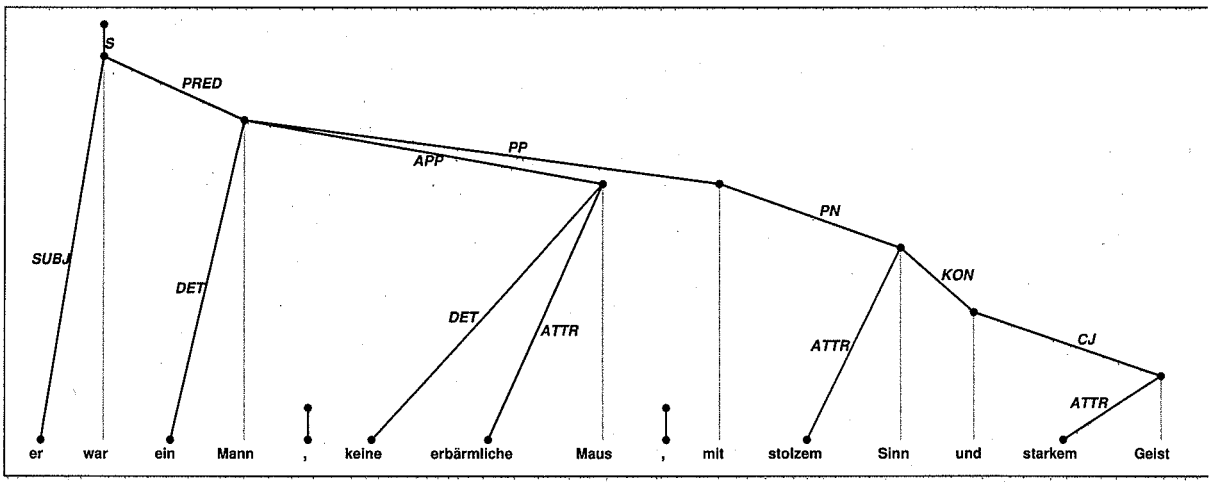
bei mehrfachem Auftreten

Reihungen von mehreren PP, insbesondere 'von'/'bis'/'über'/'auf', werden stets parallel beigeordnet, niemals untereinander.

an komplexe NP

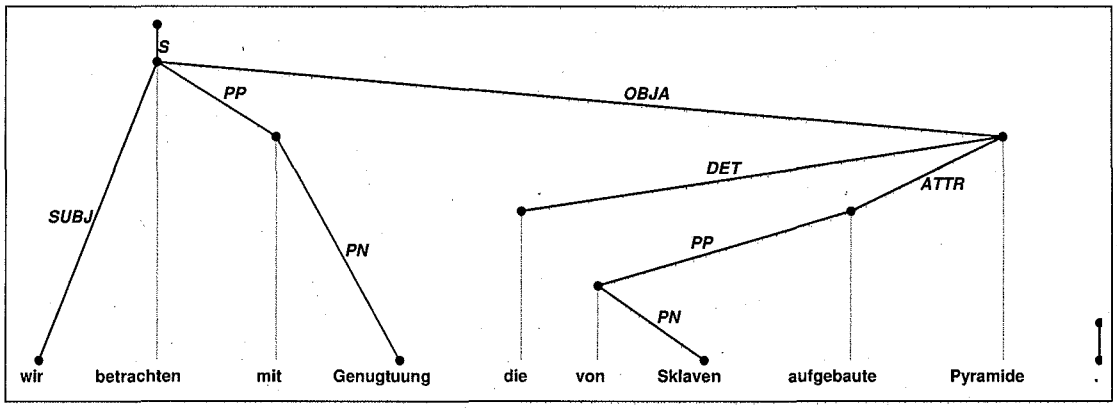
Modifiziert eine Präposition eine komplexe Nominalphrase, die aus mehreren Nomen mit APP zusammengesetzt ist, so wird sie kurz angebunden, also an das letzte Nomen. Nur wenn der Sinn diese Möglichkeit eindeutig ausschließt, wird lang angebunden.



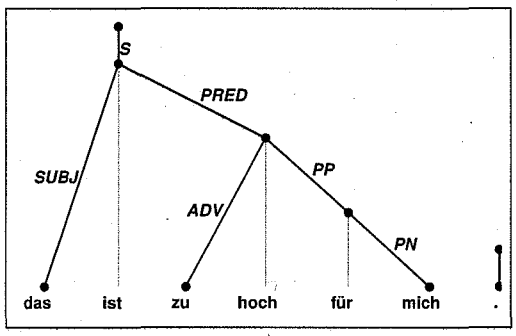


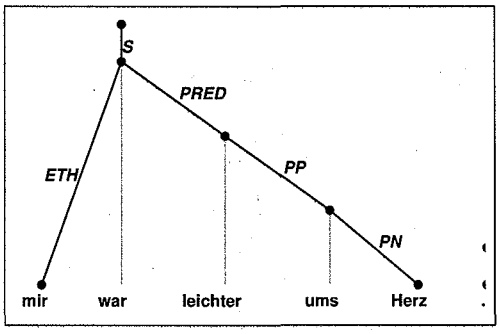
an Adjektive

Präpositionen modifizieren nur dann Adjektive, wenn sie offensichtlich zu diesen gehören, z.B. wenn sie in die NP eingeschachtelt sind.

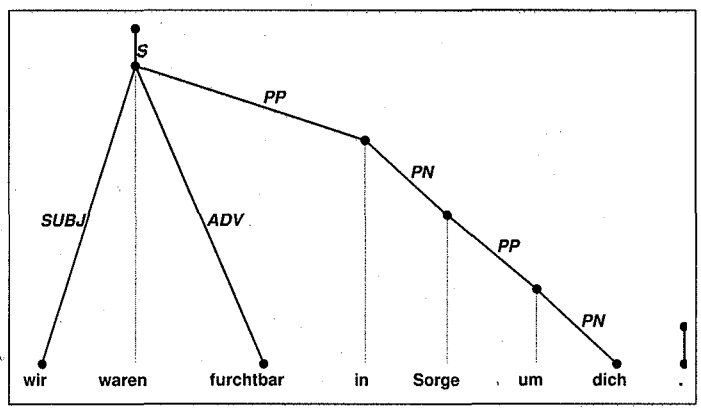


Wenn die Präposition einem prädikativ gebrauchten Adjektiv folgt (Label PRED), ist sie meistens als dazugehörig anzusehen.

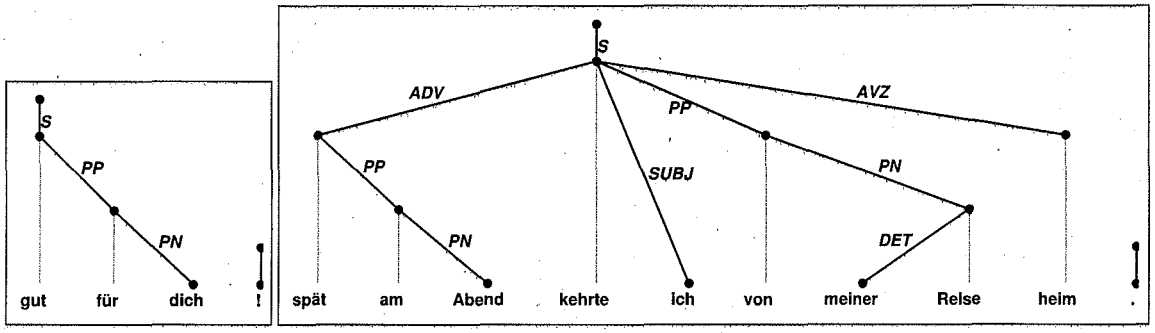




Wenn Präposition und prädikatives Adjektiv sonst nebeneinander auftreten, sollen sie normalerweise beide denselben Regenten haben.



Wenn das nicht möglich ist, etwa im Fragment oder im Vorfeld, soll die Präposition das Adjektiv modifizieren und nicht umgekehrt:

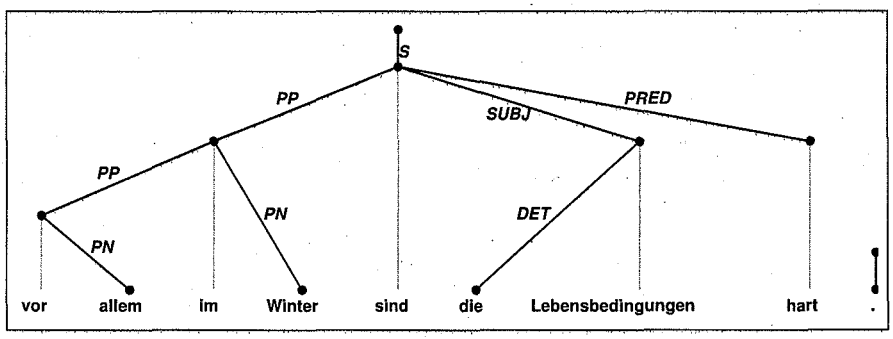


innerhalb von Verbgruppen

Hier gelten dieselben Regeln wie für Adverben (vgl. 'Anbindung von ADV', Seite 85).

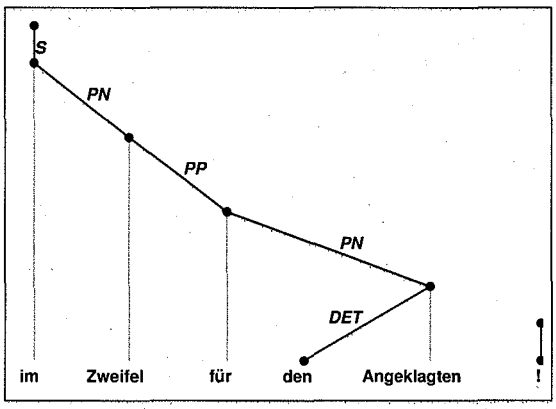
an Präpositionen

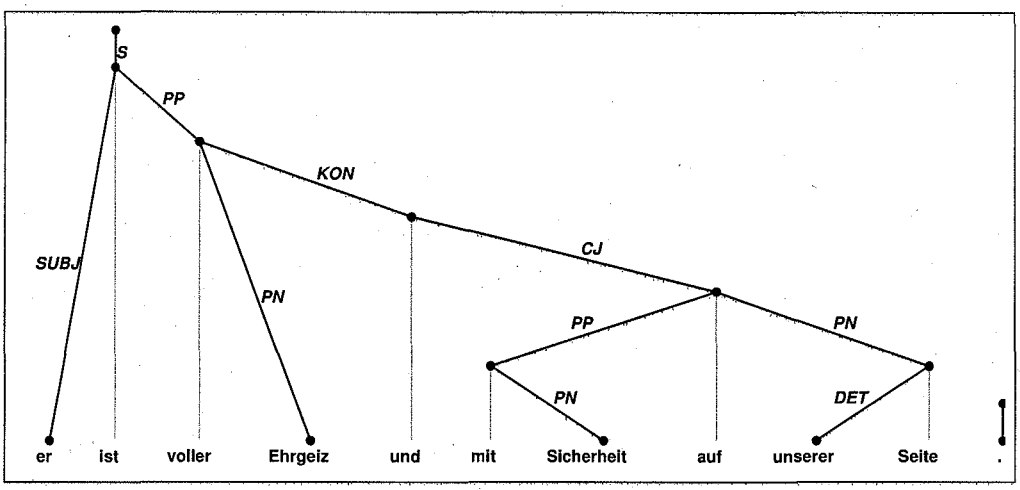
Präpositionen modifizieren nur dann andere Präpositionen, wenn sie verblaßte Adjektive tragen, die den Charakter von Adverben haben. Anderenfalls sind sie nebenzuordnen, auch wenn dadurch z.B. das Vorfeld überfüllt wird.



- *Im Winter in Sibirien sind die Lebensbedingungen hart.

Präpositionen dürfen außerdem dann andere Präpositionen modifizieren, wenn diese fragmentarisch oder beigeordnet sind.





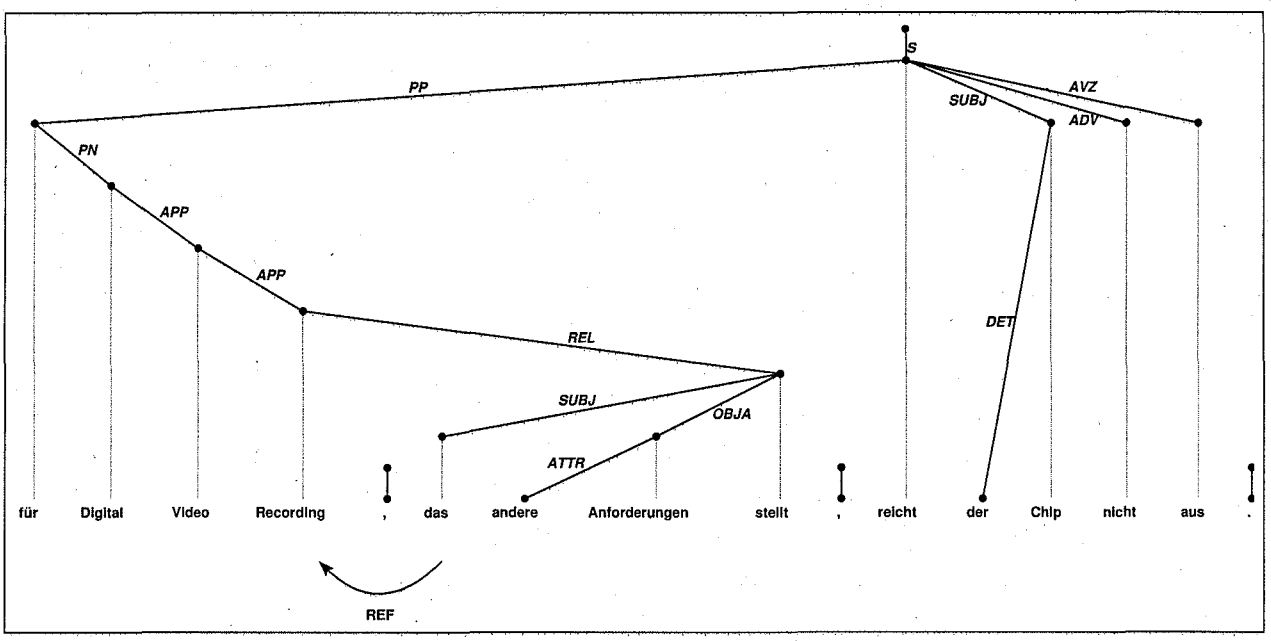
Anbindung von REF

Kanten auf der Ebene REF ordnen nur Relativpronomen unter. Sie modifizieren jeweils dasjenige Wort, das den betreffenden Relativsatz trägt. Wenn mehrere Relativsätze beigeordnet sind, modifizieren beide Relativpronomen dasselbe Bezugswort.

Andere Pronomen werden nicht durch REF untergeordnet, auch wenn das Bezugswort im selben Satz zu finden wäre.

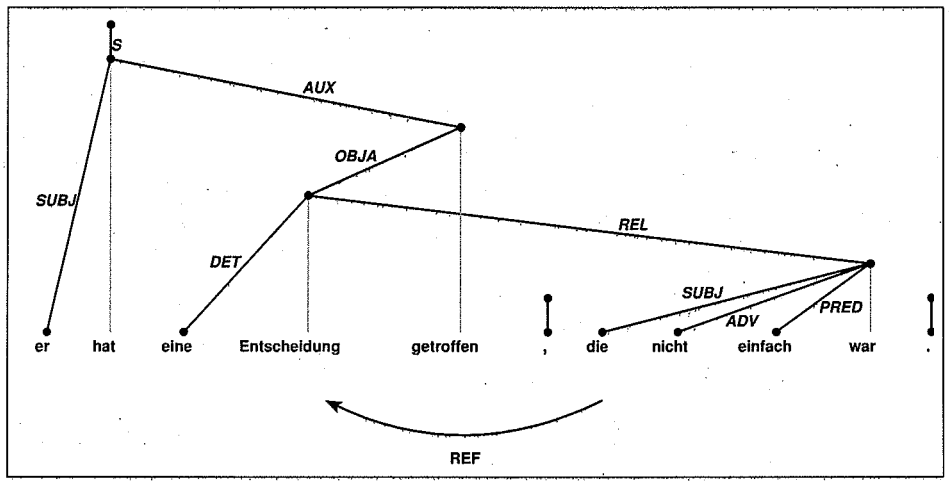
Anbindung von REL

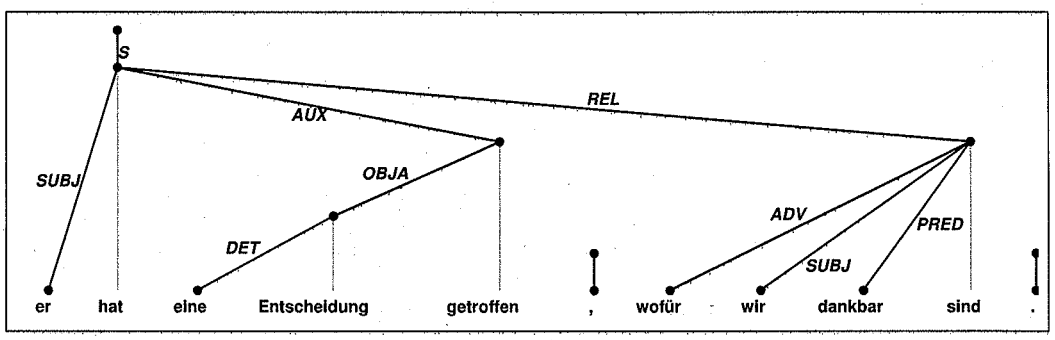
Relativsätze modifizieren gewöhnlich das Nomen, das der Referent ist. In mehrteiligen NP wird das nächste Nomen modifiziert, also gewöhnlich das letzte.



Relativsätze, die ganze Sätze erläutern, modifizieren das finite Verb.

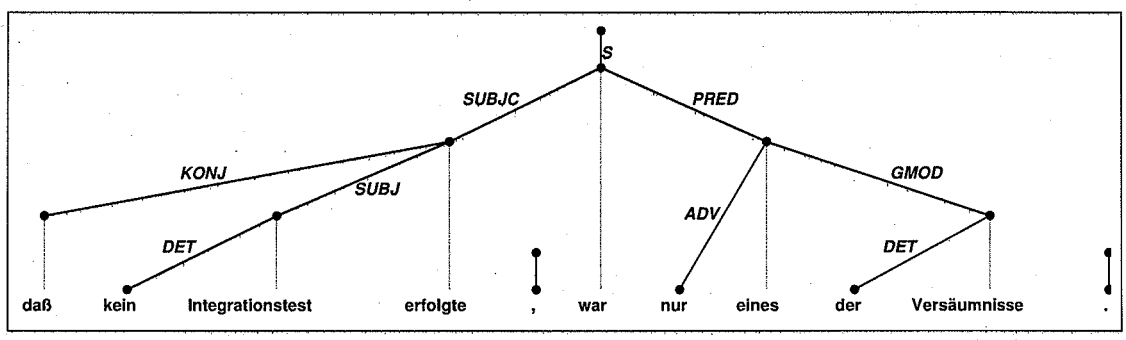
Ist unklar, ob NP oder VP der Referent ist, so modifizieren die Relativpronomen der 'das'-Reihe eher Nomen und die der 'was'-Reihe eher Verben.



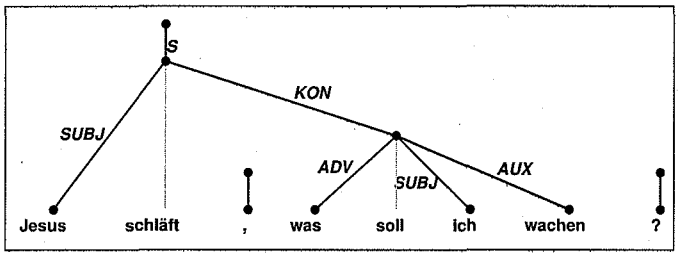


Anbindung von SUBJ

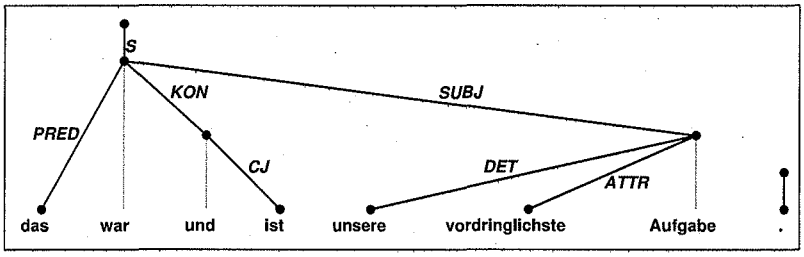
Das Subjekt steht stets an demjenigen Verb, zu dem es inhaltlich gehört.



Sind mehrere Sätze koordiniert, so modifiziert jedes Subjekt das Verb seines Teilsatzes.

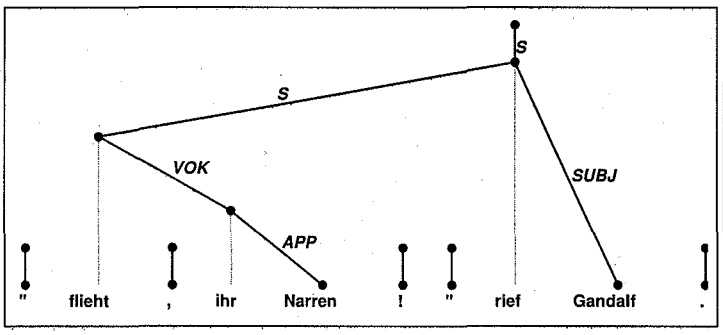


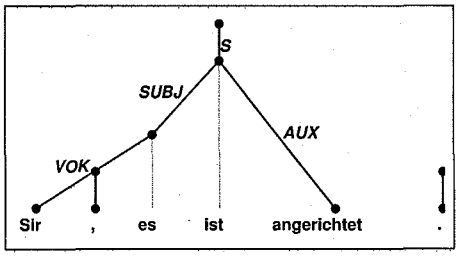
Sind zwei Verben eng koordiniert, d.h. ohne zweites Subjekt oder Objekt, dann modifiziert das Subjekt das übergeordnete Verb.



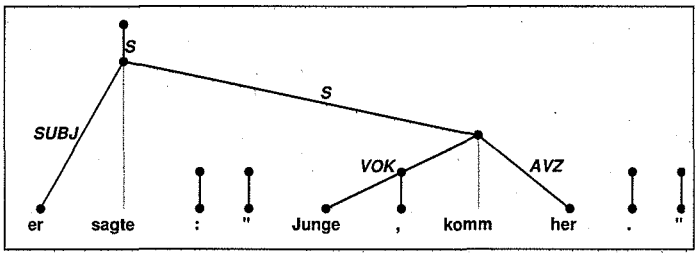
Anbindung von VOK

Die Anrede modifiziert das nächste Wort (nicht Satzzeichen) links von ihr. Steht sie ganz vorn, so modifiziert sie stattdessen das nächste Wort rechts.



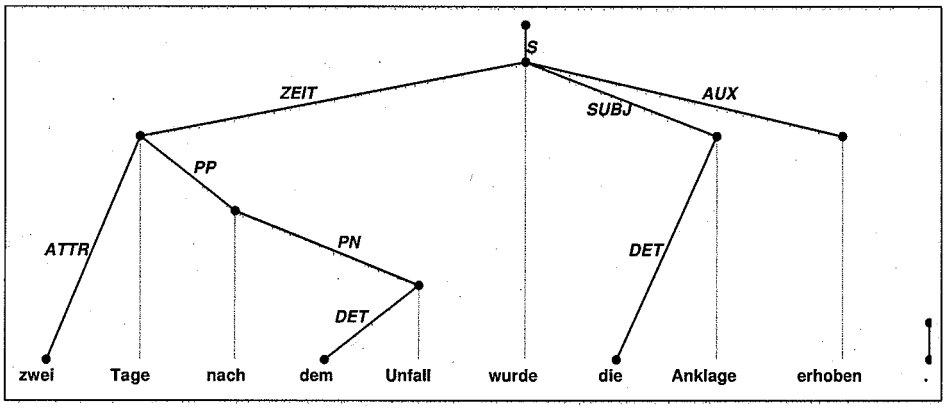


Wenn sie als erstes in einer direkten Rede steht, so modifiziert sie das nächste Wort, auch wenn der Matrixsatz vorausgeht.



Anbindung von ZEIT

Wenn eine absolute Zeitangabe zusammen mit einer Präposition auftritt, modifiziert die Zeitangabe das Verb und die Präposition die Zeitangabe.

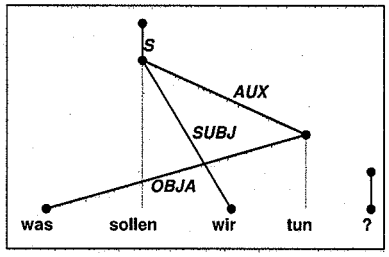


Projektivitätsregeln

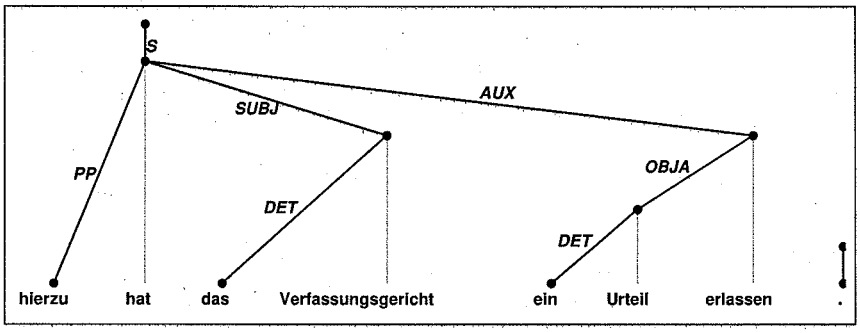
Grundsätzlich sind nur projektive Syntaxstrukturen erlaubt. Ausnahmen werden nur für eine bestimme Gruppe von Phänomenen gemacht, die alle mit der Verbletztstellung der Klammerstellung zusammenhängen.

1. Komplexe Verbphrasen

   (a) mit Topikalisierung:

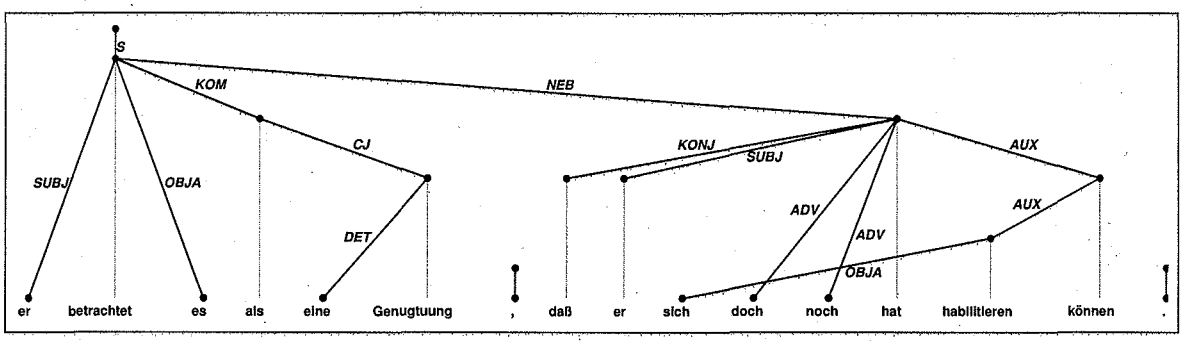


Hier kollidieren die Regeln für Verbgruppen (VVINF unter VMFIN) mit der Regel für Vollverben (OBJA nur an Vollverb). Statt eine von beiden aufzugeben, erlauben wir die nichtprojektive Anbindung des Objektes. Das gilt auch für OBJD, OBJG, etc. Insbesondere gilt es für OBJP, aber **nicht** für PP oder KOM:



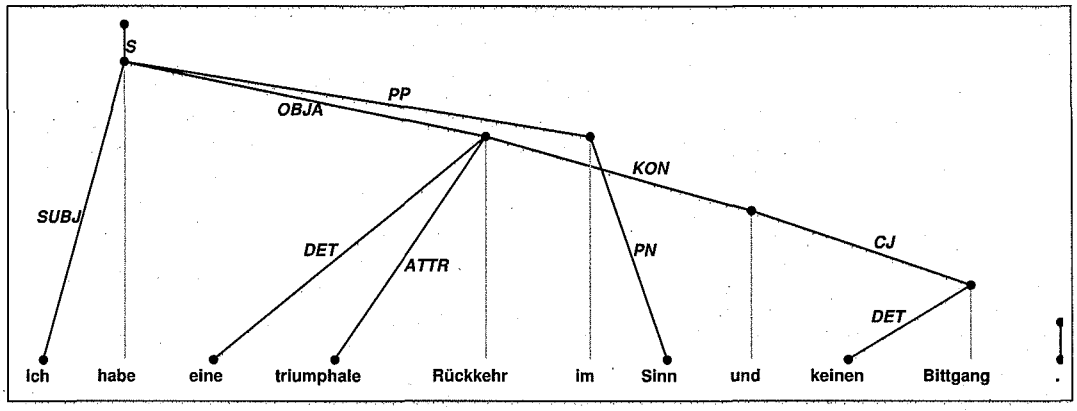
Hier muß die Präposition das finite Verb modifizieren, obwohl inhaltlich die Alternative etwas passender wäre.

   (b) mit Inversion:



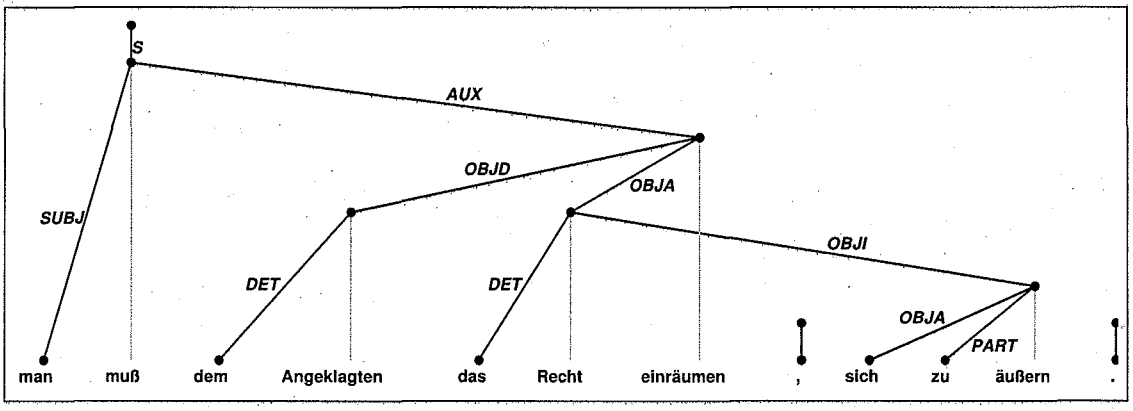
2. Koordinationen und Appositionen

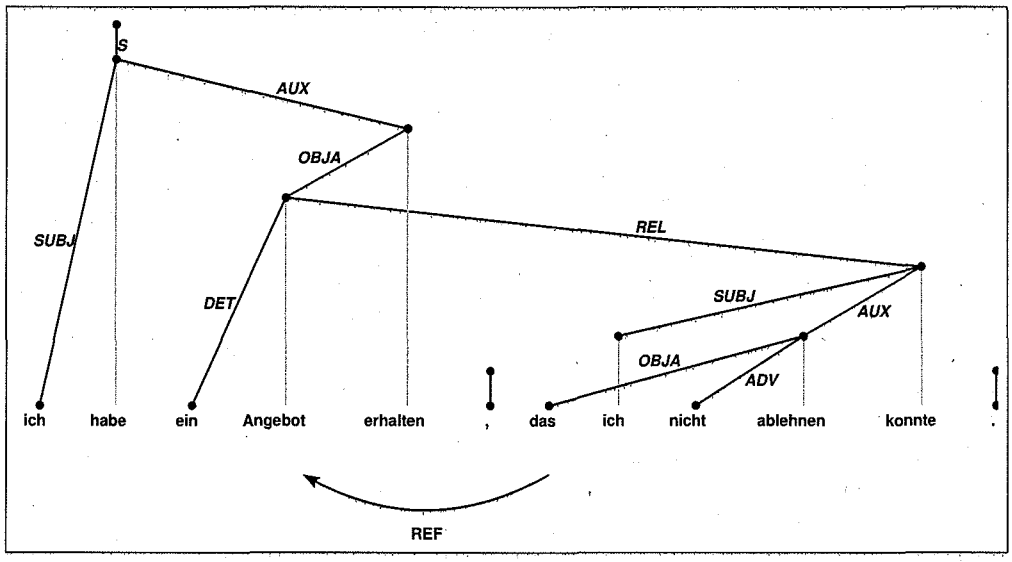
Sowohl markierte als auch unmarkierte Koordinationen können nach rechts extraponiert werden, auch wenn dadurch die Projektivität verletzt wird:



3. Relativsätze und Objektsätze

Diese Art von untergeordneten Sätzen ist fast immer extraponiert. Nichtprojektivität kommt zustande, wenn das Bezugswort in eine komplexe Verbphrase verschachtelt ist.





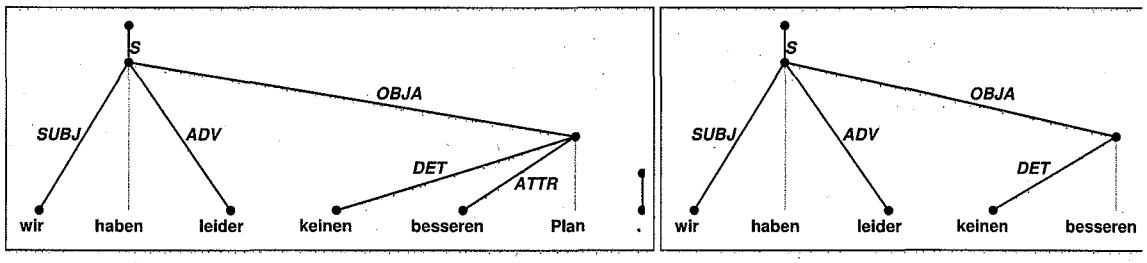
Ellipsen

Das schwerste und häufigste Repräsentationsproblem stellen Ellipsen dar. Fast jede Konstituente eines regelhaften Satzes kann fortgelassen werden. Insbesondere kann der Kopf einer Phrase fehlen, obwohl andere Teile der Phrase verbleiben. Will man dem Parser nicht erlauben, spontan zusätzliche virtuelle Anbindungen zu erlauben, so müssen diese Worte als Fragment behandelt werden.

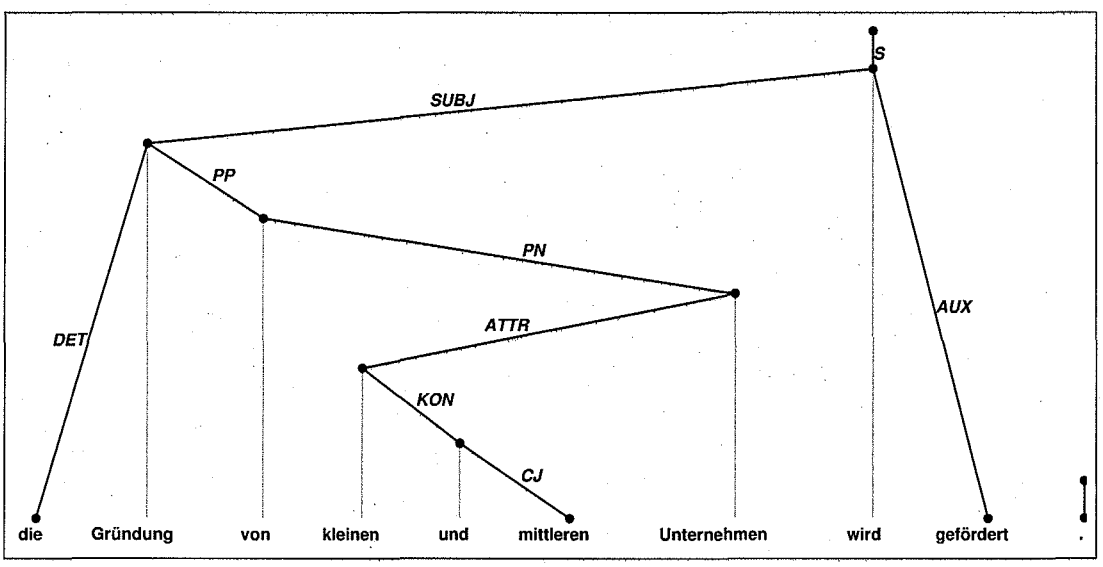
In einigen Fällen dürfen die Kinder von fortgefallenen Worten ersatzweise deren Regent modifizieren, in anderen Fällen müssen sie zum Fragment erklärt werden. In vielen Fällen hängt die Unterordnung davon ab, welches der ursprüngliche Satz war. Dieser sollte jeweils soweit aus dem Zusammenhang ersichtlich rekonstruiert werden. In den folgenden Beispielen wird stets davon ausgegangen, daß der Kontext den eingeklammerten Satz als Ursprungsversion nahelegt.

- von Nomen

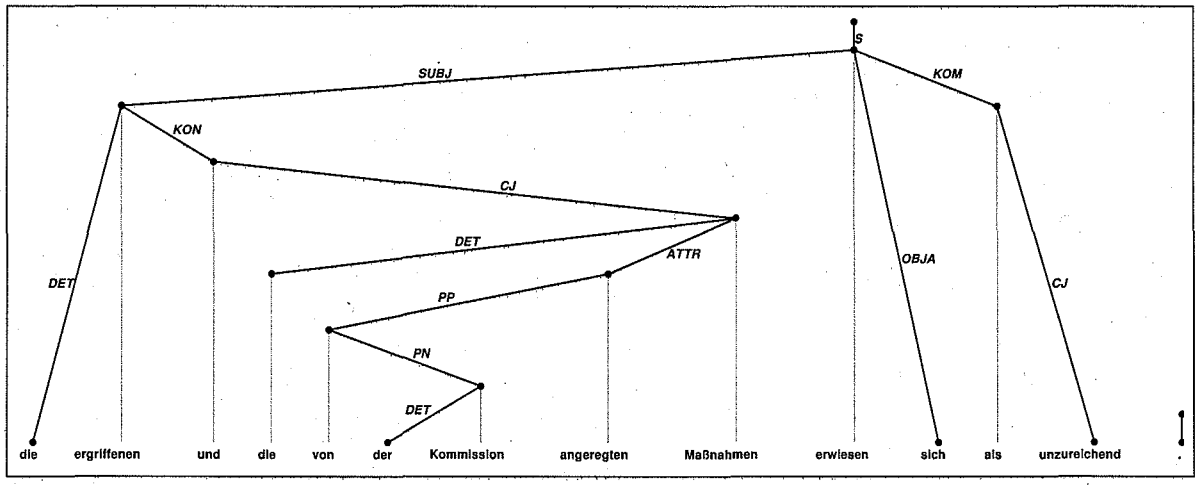
  Fällt ein Nomen weg, so kann ein attributives Adjektiv dessen Rolle einnehmen. Es kann SUBJ, OBJA etc. sein und durch Artikel als DET modifiziert werden. Ist kein Adjektiv vorhanden, so sind evtl. verbleibende Vorkommen von 'der' etc. immer Demonstrativpronomen, nicht Artikel.



  Besonders häufig fällt ein Nomen fort, wenn zwei NP mit demselben Kopf koordiniert werden. Oft kann diese Konstruktion als koordinierte Adjektivphrase modelliert werden:



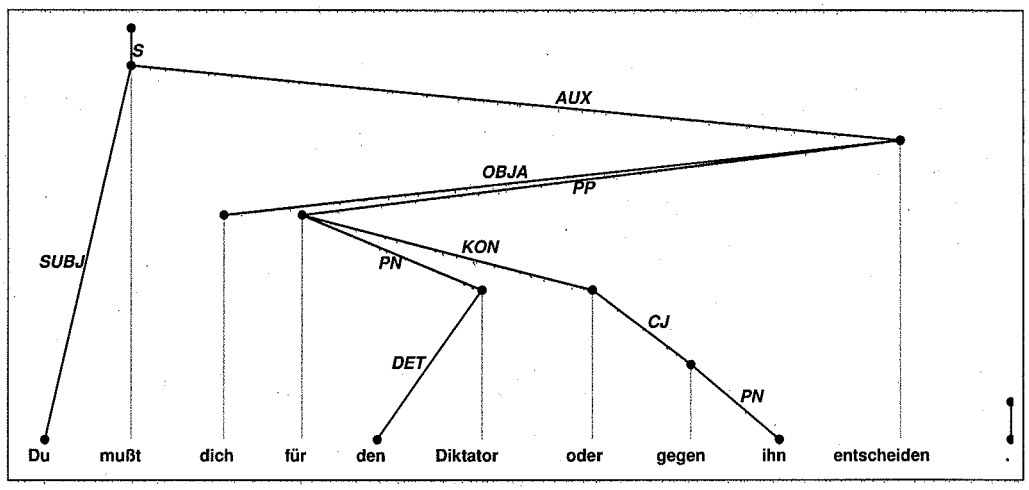
Wenn von der ersten NP mehr verbleibt als ein bloßes ADJA, ist meist eine koordinierte NP anzunehmen. In diesem Fall ist das zweite Nomen als Koordination zum verbliebenen Adjektiv anzusehen:

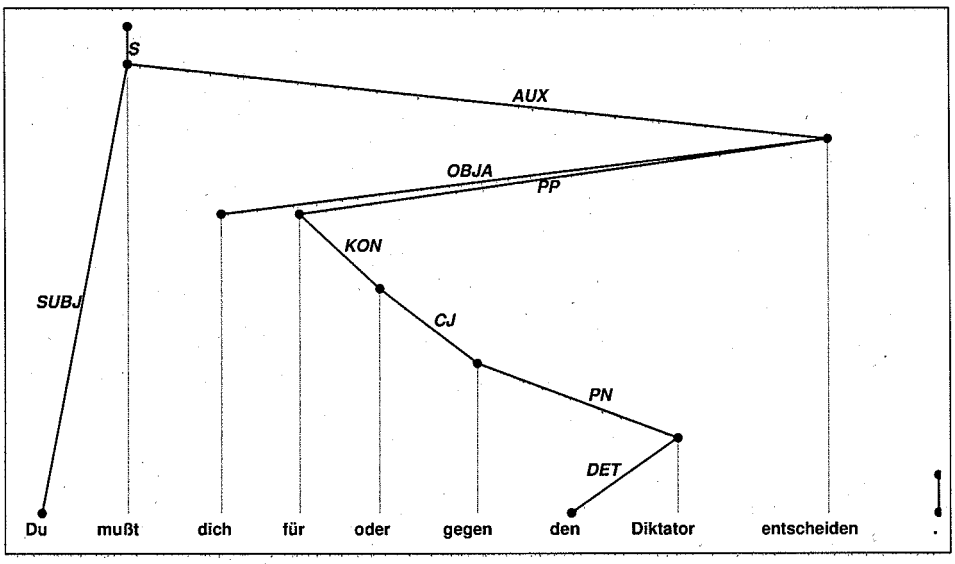


Ein Eigenname als Genitivattribut kann ebenfalls die Funktion des entfallenen Nomen einnehmen (vgl. 'GMOD oder SUBJ?').

Fällt ein Subjekt-Demonstrativpronomen weg, das einen Relativsatz trägt, so kann dieser ersatzweise als SUBJC untergeordnet werden.

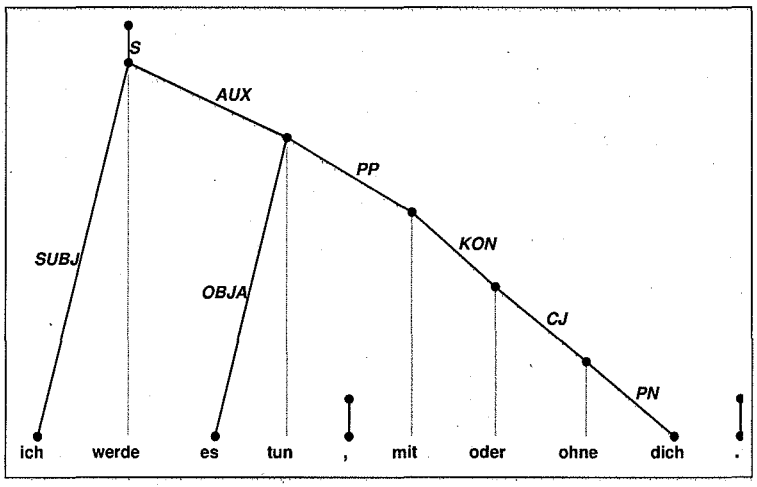
Wenn ein Nomen in zwei aufeinanderfolgenden PP auftreten würde, fällt es fast immer weg:





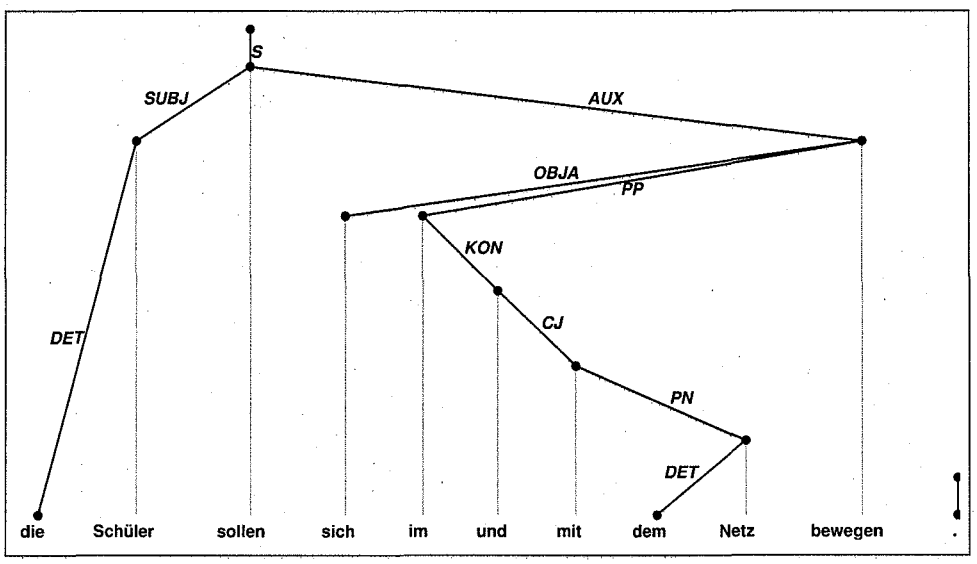
– *Du mußt dich für den Diktator oder gegen den Diktator entscheiden.

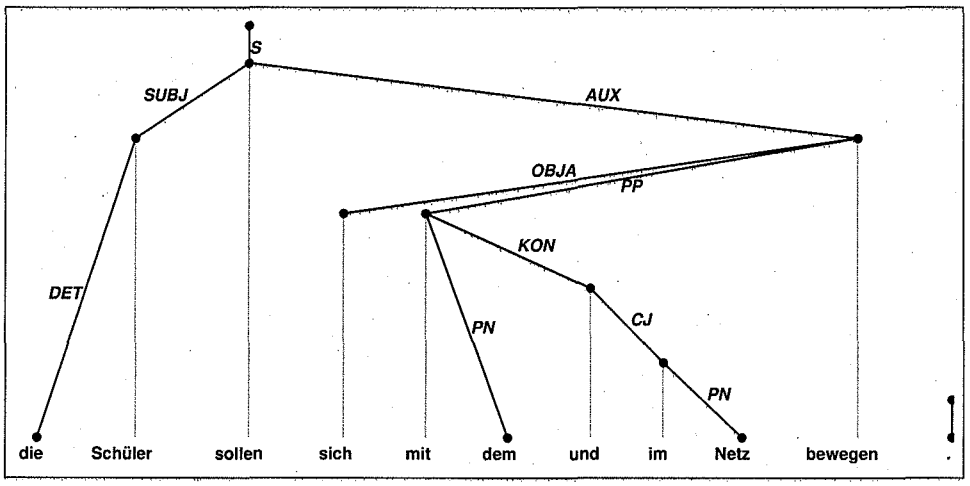
In diesem Fall kann nur eine der beiden Präpositionen ihr Argument behalten. Im allgemeinen scheint das Nomen eher zur zweiten Präposition zu gehören; wenn zum Beispiel die Präpositionen unterschiedliche Kasus nehmen, gehorcht es eher der zweiten Präposition:



– *Ich werde es tun, mit oder ohne dir.

Auch wenn Artikel im Spiel sind, scheint das Nomen mit der zweiten Präposition eine Einheit zu bilden:



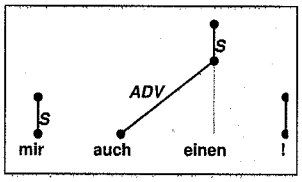


- *Die Schüler sollen sich im und mit Netz bewegen.
- *Die Schüler sollen sich mit und im dem Netz bewegen.

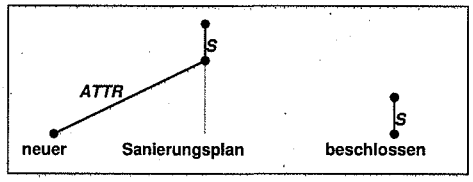
Deshalb soll das Nomen immer kurz, d.h. die zweite Präposition, modifizieren.

- von Verben als Satzwurzel

  Fällt ein finites Verb weg, so werden Subjekt und Objekte zu Fragmenten. Es gibt **keine** Unterordnung zwischen ihnen.

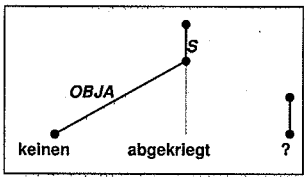


(Gib mir auch einen!)

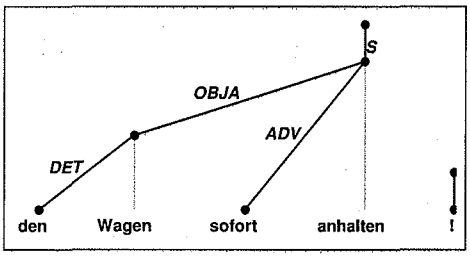


(Ein neuer Sanierungsplan ist beschlossen worden.)

Verbleibt das Vollverb im Satz, so kann es natürlich wie üblich Objekte und adverbiale Bestimmungen tragen. Insbesondere bei isolierten Partizipien ist darauf zu achten, ob die verbleibende NP als Subjekt oder als Objekt aufzufassen ist. Ist die NP eindeutig als Akkusativ gebeugt, so ist sie OBJA statt S.



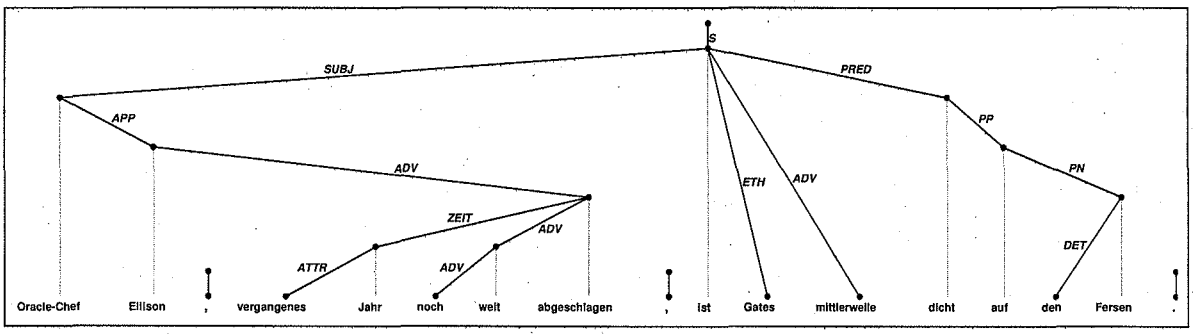
(Sie haben keinen abgekriegt?)



(Sie werden den Wagen sofort anhalten, oder wir werden Sie unter Feuer nehmen!)

- von Verben im Relativsatz

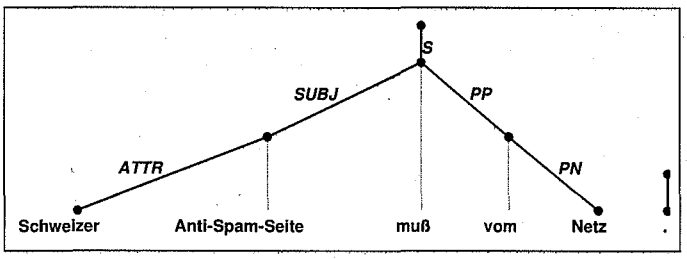
  Fällt das Hilfverb eines Relativsatzes weg, so kann ein verbleibendes Partizip ersatzweise den Referenten als ADV modifizieren.



(Oracle-Chef Ellison, der vergangenes Jahr noch weit abgeschlagen war, ist Gates mittlerweile dicht auf den Fersen.)

- von VVINF

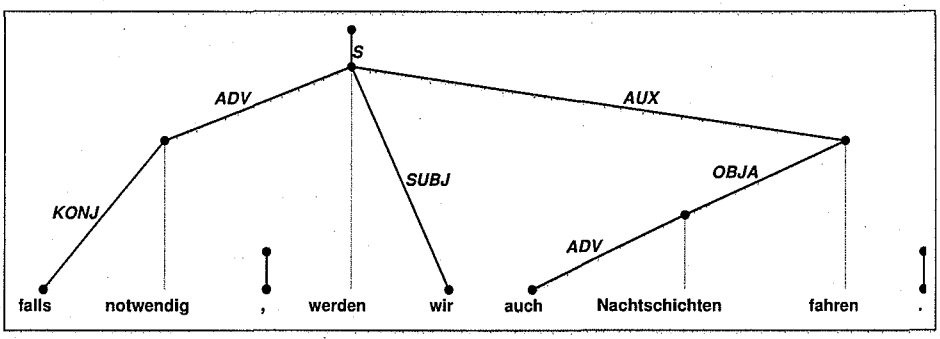
  Fällt ein infinites Vollverb weg, so modifizieren alle adverbialen Bestimmungen (ADV, PP, KOM) stattdessen das Hilfsverb. Ihre Label bleiben unverändert, d.h. die fehlende AUX-Kante wird nicht ersetzt.



(Schweizer Anti-Spam-Seite muß vom Netz verschwinden.)

- von prädikativen Nebensatzverben

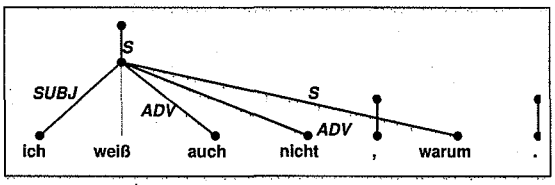
  In Fällen wie 'falls nötig' können Konjunktionen ersatzweise das Adjektiv als KONJ modifizieren. Das Adjektiv selbst ist dann ADV.



(Falls es notwendig wird, werden wir auch Nachtschichten fahren.)

- in indirekten Fragesätzen

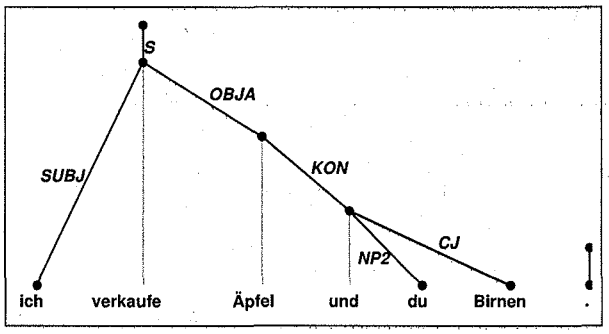
  Wird ein Relativsatz auf das Fragewort verkürzt, so darf dieses als S dem Hauptsatz untergeordnet werden:



(Ich weiß auch nicht, warum er das getan hat.)

- von koordinierten Verben

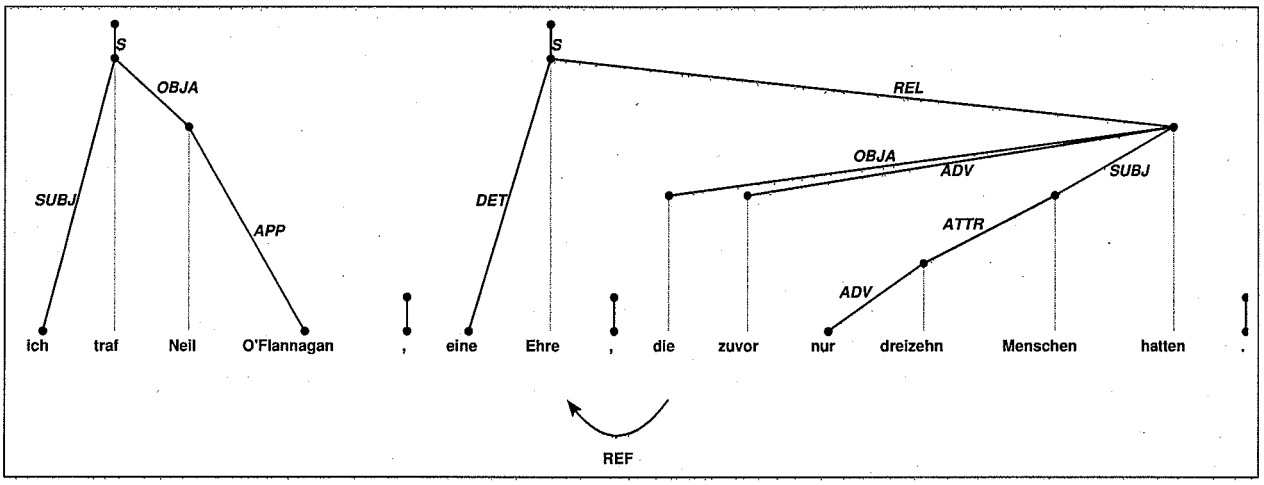
  Sehr oft fällt die wörtliche Wiederholung eines Verbs fort, wenn mehrere ähnliche Sätze beigeordnet werden. Dadurch wird eines der beigeordneten Worte heimatlos und kann nur als Fragment analysiert werden. Meistens ist das das Subjekt des zweiten Satzes. In solchen Fällen kann hier das Label `NP2` verwendet werden.



(Ich verkaufe Äpfel und du verkaufst Birnen.)

- von Verb und Konjunktion gleichzeitig

  Oft wird eine NP alleinstehend verwendet, wenn sie dazu dient, die vorige Aussage zusammenzufassen. In diesem Fall ist die alleinstehende NP `S`.



(Ich traf Neil O'Flannagan, und das war eine Ehre, die zuvor nur dreizehn Menschen hatten.)

1.2.6 Welche morphologische Variante?

nom oder acc?

Wenn die syntaktische Funktion von Fragmenten eindeutig zu erkennen ist, ist die Variante zu wählen, die im ganzen Satz gestanden hätte.

- Hände/acc hoch! (= Nehmt die Hände/OBJA hoch!)
- Mann/nom über Bord! (= Ein Mann ist über Bord gefallen!)

Wenn die ursprüngliche Funktion nicht zu rekonstruieren und die Form mehrdeutig ist, ist vorzugsweise der Nominativ oder der Singular zu wählen.

- Piraten/nom!
- Streber/sg!

Homonyme und 'subcat'

Das feature 'subcat' trennt Homonyme, die sich nicht in ihrer STTS-Kategorie unterscheiden, aber dennoch syntaktisch unterschiedlich verhalten.

Bei Eigennamen wird beispielsweise eine Unterscheidung zwischen Vornamen, Nachnamen, geographischen Namen, Institutionen und Produkten getroffen. Es ist jeweils die Variante anzunehmen, die aus dem Gebrauch zu erschließen ist, soweit möglich.

- Auf dem Programm stehen Werke von Thomas/Vorname Tallis und Ambroise Thomas/Nachname.
- Israel/Vorname Fleming wohnte lange Zeit in Israel/Region.

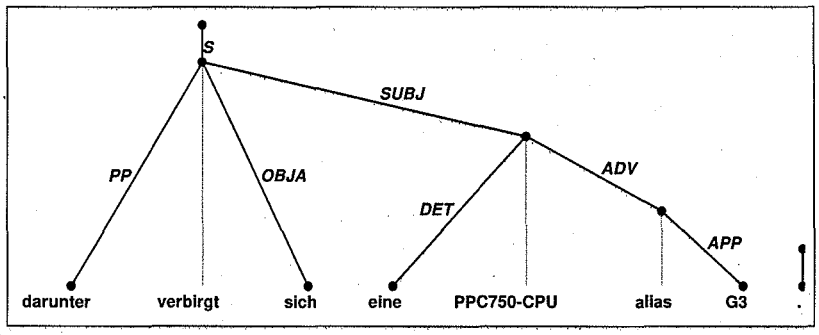
Einige Adverben können entweder Zeitsinn oder aber übertragene Bedeutung annehmen und sich dabei syntaktisch unterschiedlich verhalten. Wenn der Gebrauch den Zeitsinn nahelegt, ist also 'temporal' anzunehmen, sonst 'focus'.

- Ich habe ihn noch/temporal gar nicht gefragt.
- Wir sind immer noch/temporal ganz am Anfang.
- Noch/focus in Texas hörte man die Explosion.

1.2.7 Einzelne Konstruktionen

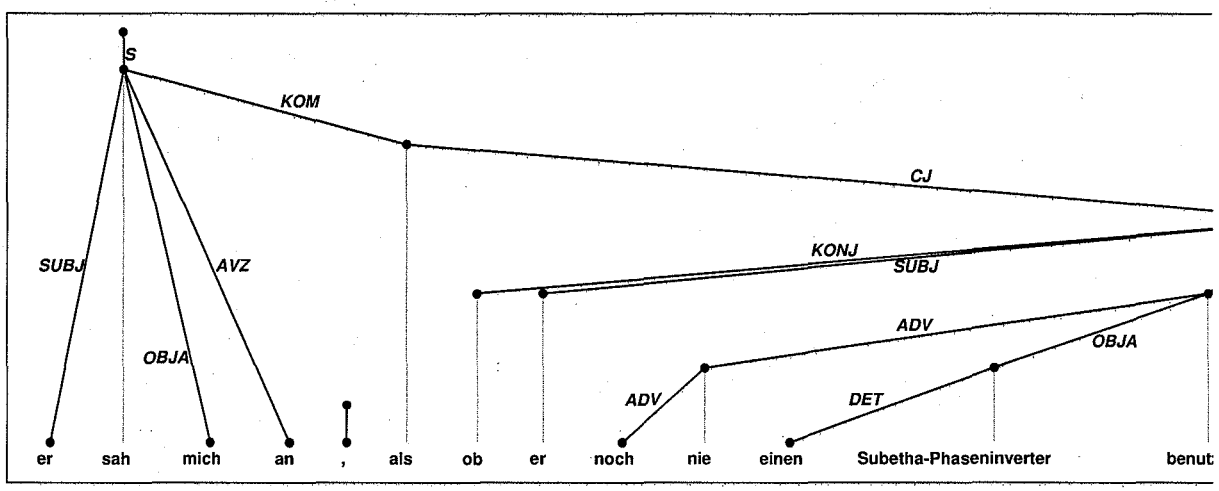
'alias'

Das Fremdwort 'alias' ist sinngemäß als Adverb anzusehen, verhält sich gleichzeitig aber auch wie ein Nomen. Zwischen zwei Namen ist es als ADV zu bezeichnen, während der zweite Name APP zum 'alias' ist:



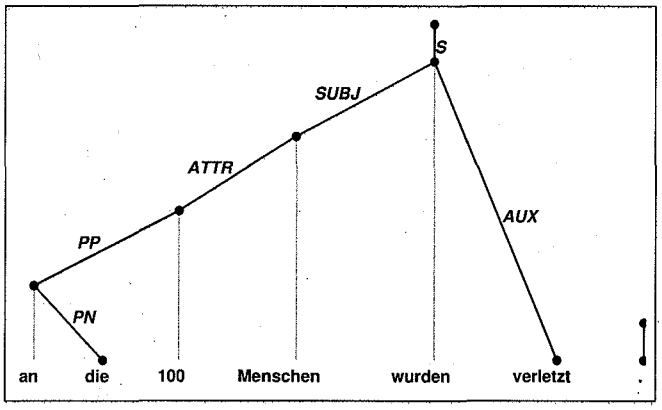
'als ob'

Die Fügung 'als ob' gilt als Vergleich eines Satzes mit einem anderen. Folglich ist 'als' KOKOM, und das zweite Verb modifiziert das Wort 'als' als CJ.



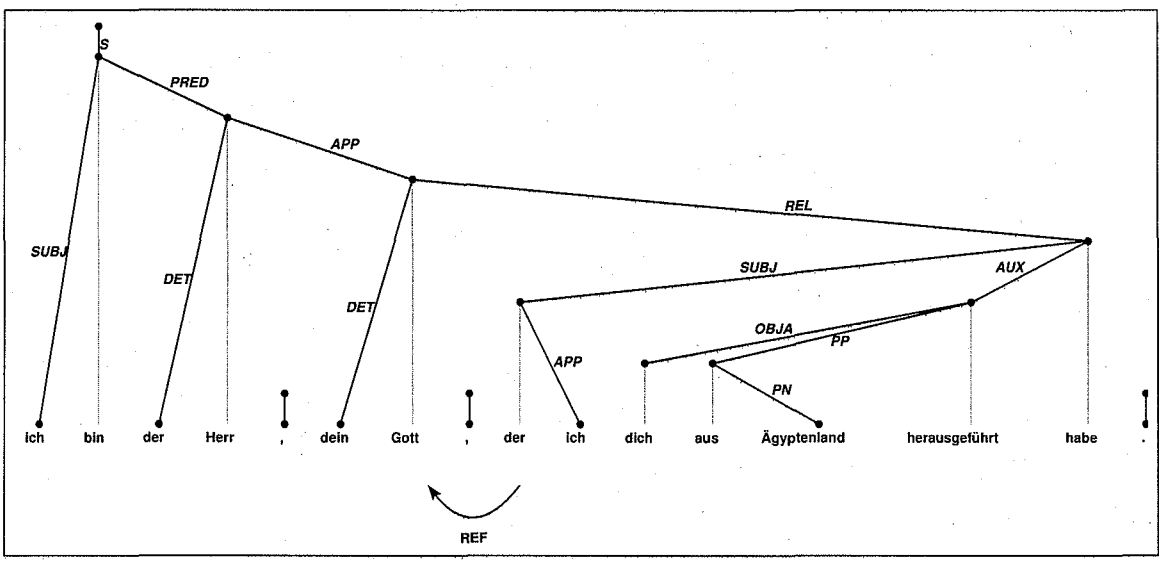
'an die'

Die Fügung 'an die' im Sinne von 'etwa' wird als Paar von Präposition+Artikel behandelt und normal als PP untergeordnet, und zwar der Zahl, die sie modifiziert. (Nur in dieser Konstruktion darf ein Artikel das Label PN tragen.)



'der ich'

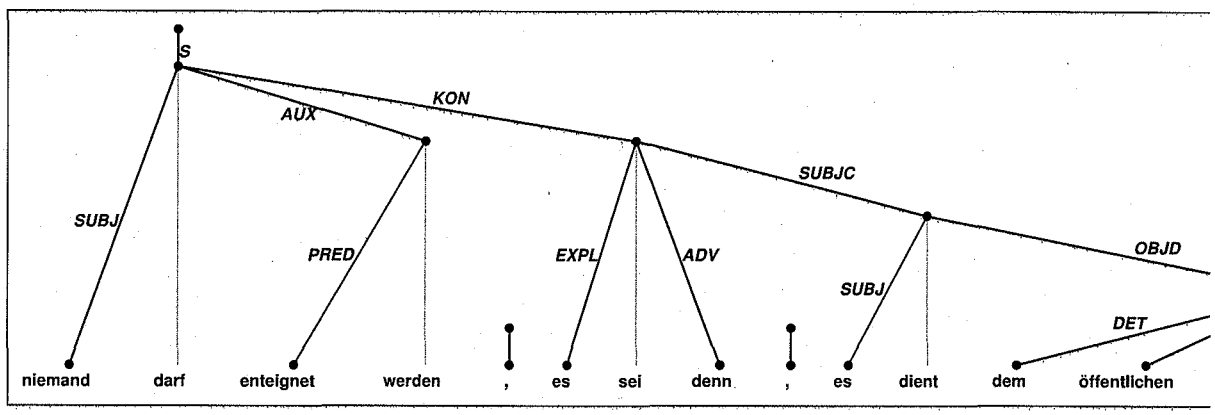
In altertümlichem Deutsch kann ein Relativsatz sowohl ein PPER als auch ein PRELS als Subjekt haben:



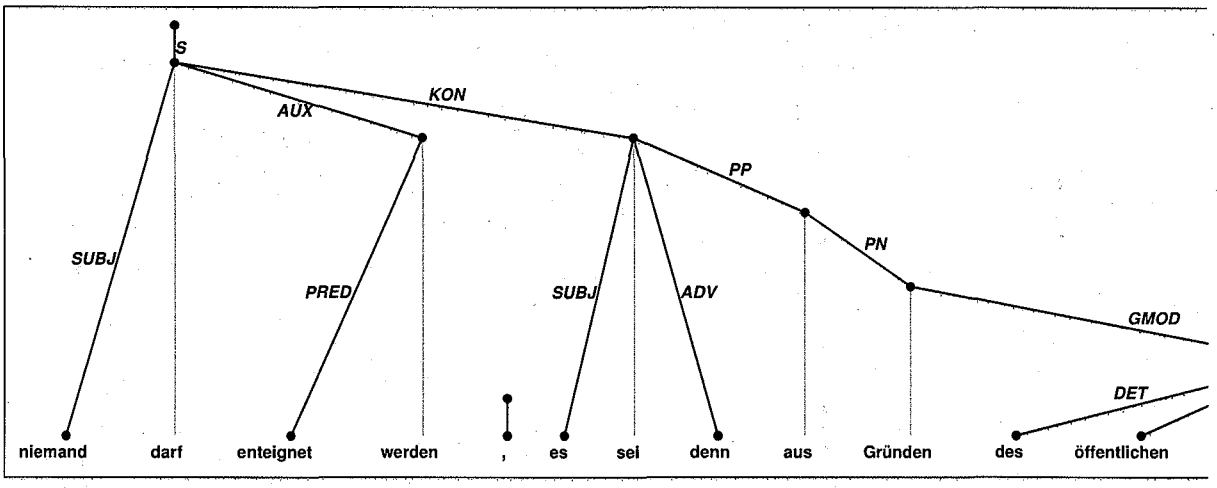
Da nicht beide Pronomen Subjekt sein können, soll in diesem Fall das Personalpronomen als Apposition zum Relativpronomen annotiert werden.

'es sei denn'

Wenn zwei Hauptsätze mit der Phrase 'es sei denn' verbunden werden, so ist eine Nebenordnung anzunehmen. Das 'es' ist als Expletivum anzusehen und der zweite Hauptsatz als Subjektsatz:

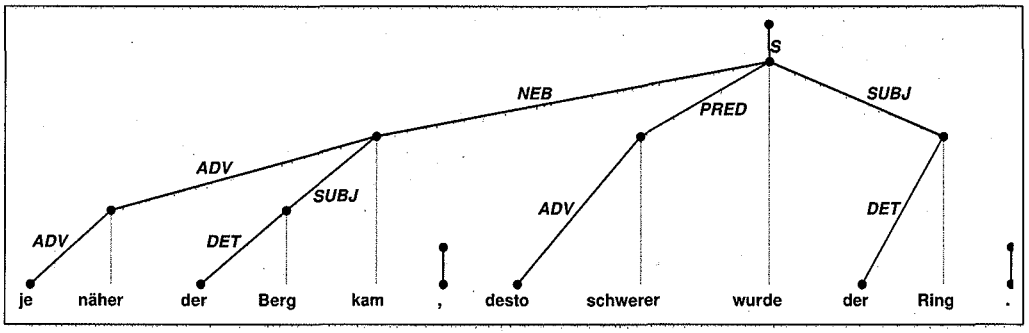


Wenn dagegen nur eine Adverbialphrase beigeordnet wird, wird sie normal in die Phrase untergeordnet:



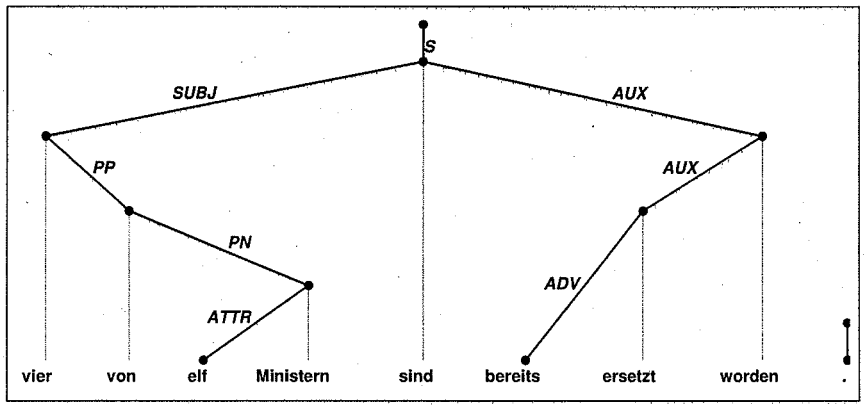
'je...desto'

Ein Nebensatz kann mit 'je' eingeleitet werden, wenn der Hauptsatz 'desto' verwendet. Beide Worte modifizieren als ADV den Komparativ, neben dem sie stehen. Das Verb des Nebensatzes modifiziert wie üblich das Verb des Hauptsatzes.



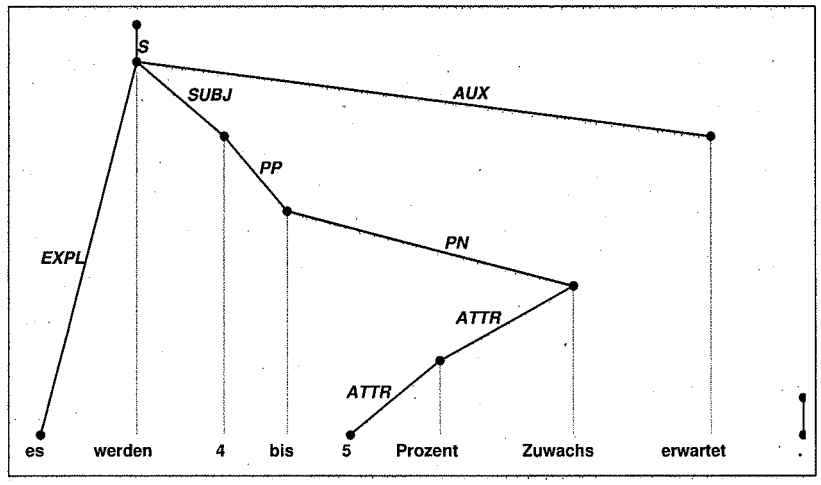
'5 bis 6 Stunden'

Wenn mehrere durch Konjunktion oder Präposition verbundene Zahlen dasselbe Nomen modifizieren, könnte sowohl die erste Zahl als auch das Nomen als Kopf der NP angesehen werden. Kongruenz und Rektion deuten jedoch darauf hin, daß das Nomen der Morphologie der PP gehorcht und nicht der externen Funktion:



- *Vier von elf Minister sind bereits ersetzt worden.
- Im neuen Modell laufen ein bis zwei G4-Prozessoren.
- *Im neuen Modell läuft ein bis zwei G4-Prozessoren.

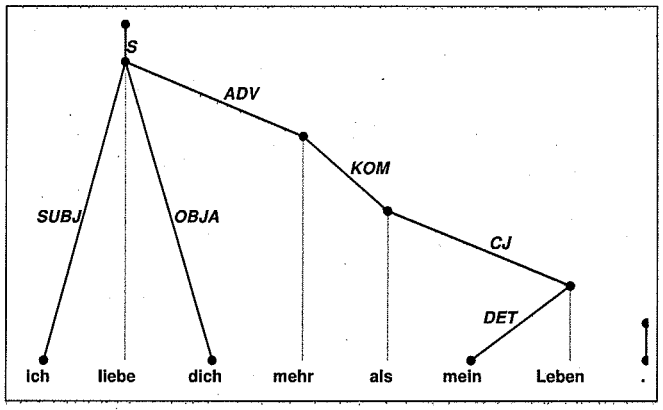
Daher wird das Nomen grundsätzlich der Präposition oder Konjunktion untergeordnet, und die zweite Zahl modifiziert das Nomen:



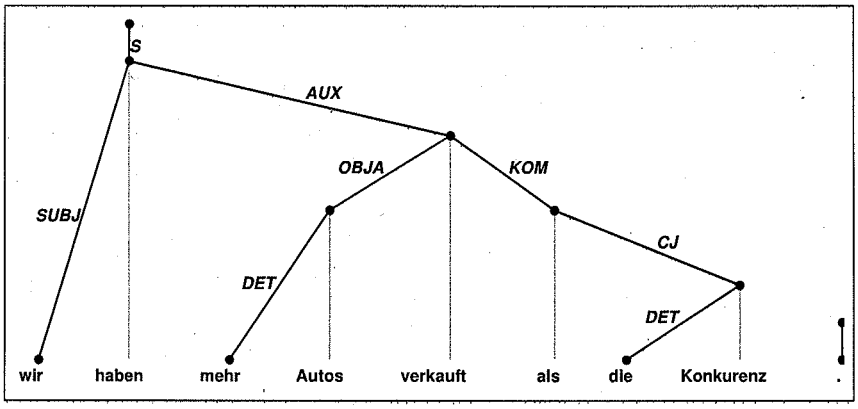
Das Wort 'ein' in der Konstruktion 'ein oder zwei' gilt dabei ebenfalls als CARD!

'mehr als'

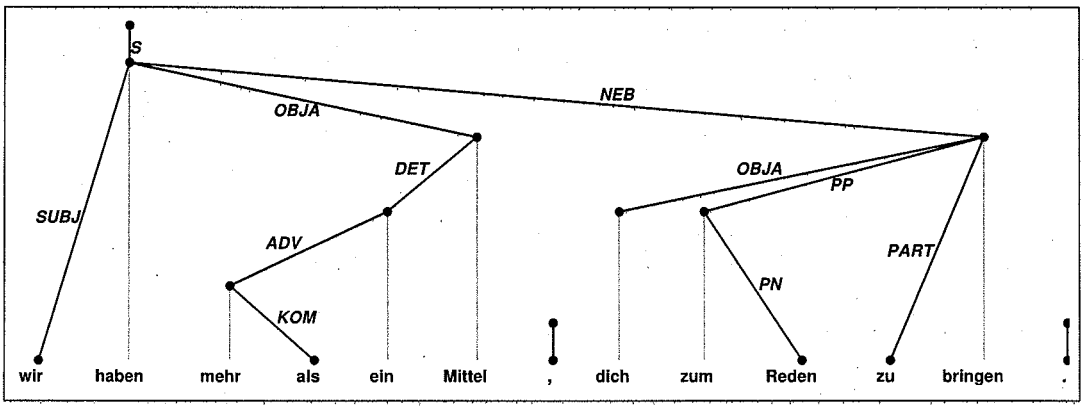
Die Ausdrücke 'mehr als' und 'weniger als' sind normale Paare von Pronomen und Vergleichswort, wenn sie adverbial vorkommen.

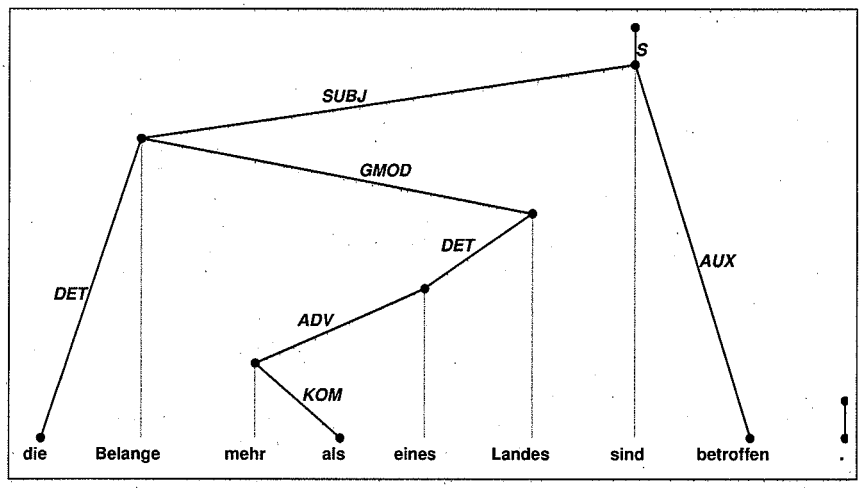


'mehr' kann auch attributiv vorkommen; auch dann ist ein späteres 'als' möglich.

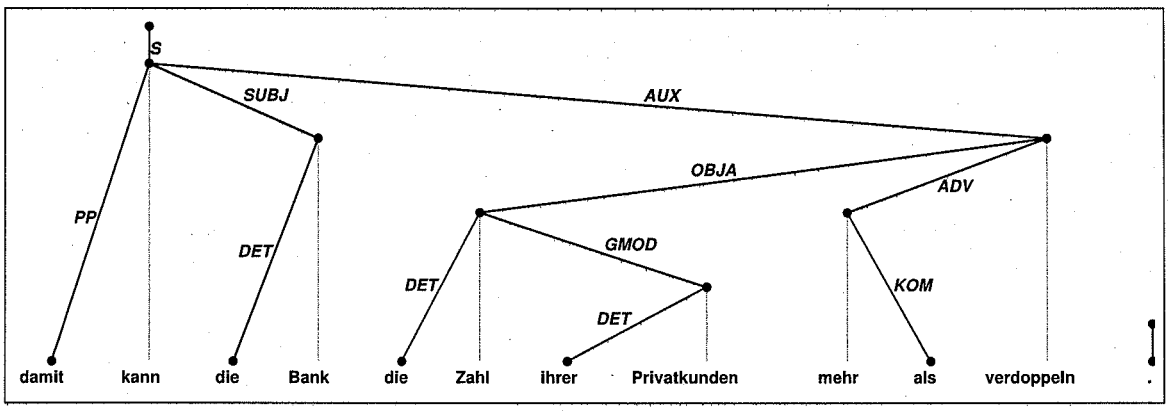


Wenn 'mehr als' aber innerhalb der Adjektivphrase auftritt, wird es behandelt wie ein mehrteiliges Adverb, d.h. 'mehr' modifiziert als Adverb das Zahlwort, und 'als' verliert sein Komplement. Dadurch ist gewährleistet, daß das Nomen der Kopf der gesamten Phrase bleibt und z.B. zur Kongruenzüberprüfung verfügbar ist.

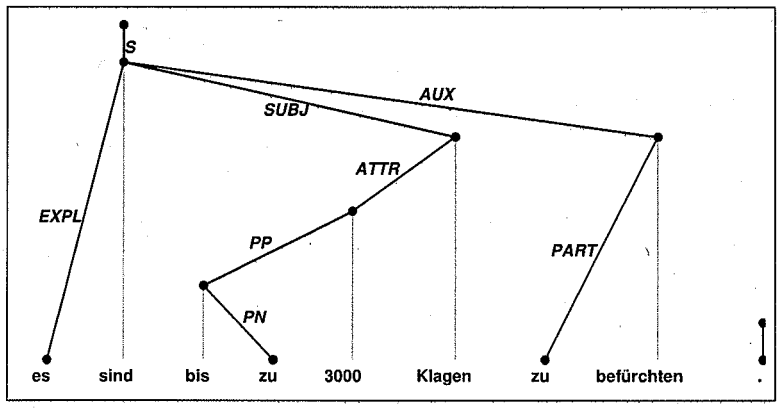




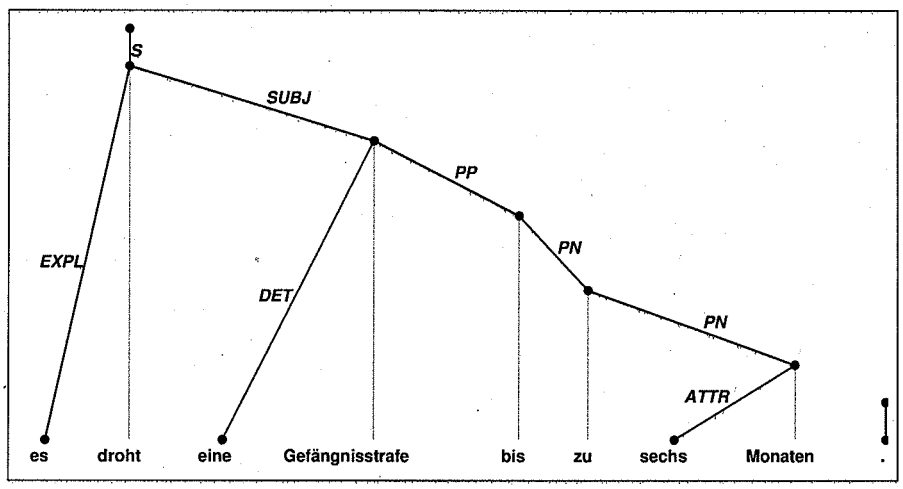
Diese Konstruktion kann auch an Verben auftreten, die Zahlsinn besitzen.



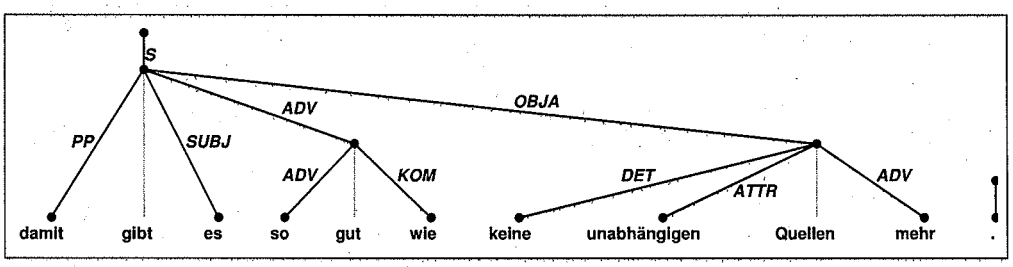
Exakt dasselbe gilt für die Fügung 'bis zu': 'bis' modifiziert das Zahlwort, und 'zu' verliert sein Argument.



Wenn aber die NP keine andere Rolle ausfüllt, ist sie normal als PN zu bezeichnen:

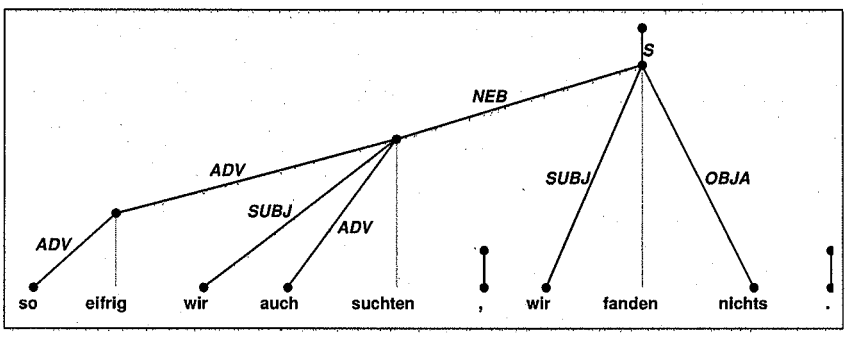


Der Quantor 'so gut wie' wird ebenfalls wie ein einzelnes Adverb behandelt, d.h. das 'wie' trägt kein Komplement:



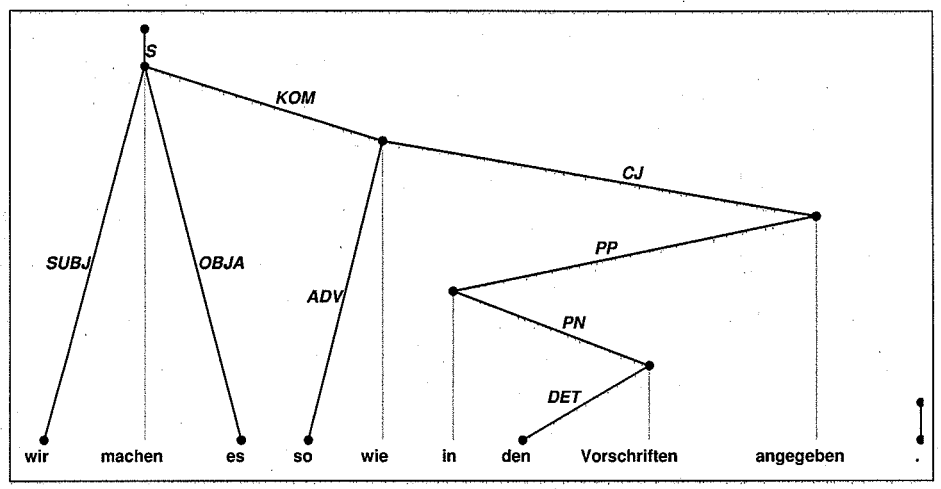
'so... auch'

Wenn ein Nebensatz mit 'so' eingeleitet wird, modifiziert 'so' als ADV das Adjektiv, vor dem es steht. Das Nebensatzverb modifiziert wie üblich das Hauptsatzverb.



'so wie'

Wenn Adverb und Vergleichswort nebeneinander treten, soll das Adverb das Vergleichswort modifizieren und nicht umgekehrt.



'um die'

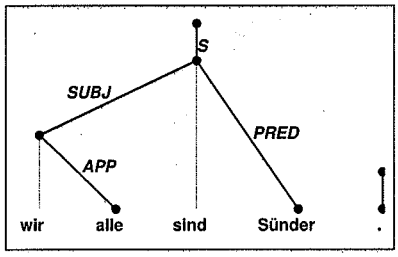
Die Fügung 'um die' im Sinne von 'etwa' wird behandelt wie 'an die' (siehe oben).

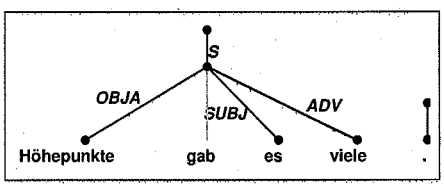
'unser aller'

Die Wendung 'unser aller [Freund, Leben, Anliegen...]' wird als Kombination aus DET und GMOD annotiert. Das word 'aller' ist als Genitiv Plural des Indefinitpronomens (PIS) 'alle' anzusehen, da es nicht mit dem Nomen kongruiert.

'wir alle'

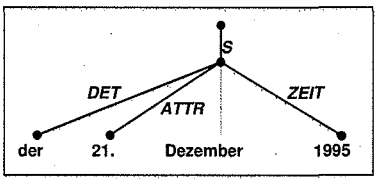
Die quantifizierenden Pronomen 'alle', 'beide' etc. können ein vorausgehendes Pronomen modifizieren, wenn sie direkt folgen. In diesem Fall sollen sie als substituierende Pronomen angesehen werden und das vorige Pronomen als APP modifizieren. Stehen sie weiter entfernt, so modifizieren sie stattdessen das finite Verb als ADV.



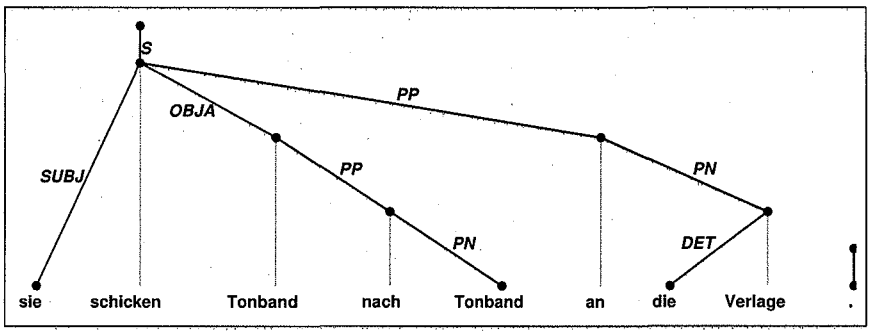


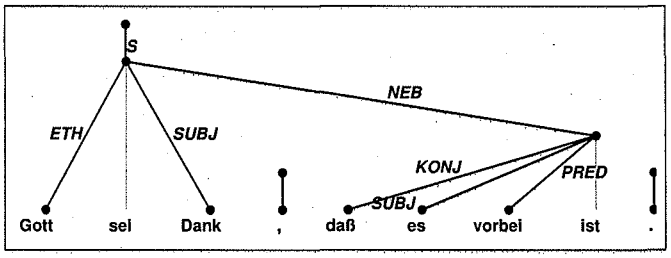
Idiomatische Ausdrücke

Datumsangaben gelten als Kombination von Attribut und Zeit-Label.

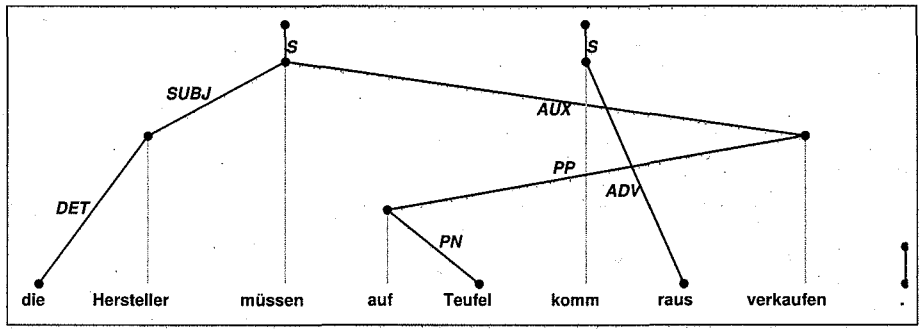


Idiomatische Ausdrücke und Konstruktionen werden soweit möglich behandelt, als wären sie freie Kombinationen.





Wenn keine reguläre Unterordnung möglich ist, müssen die betreffenden Worte als Fragmente angesehen werden.



1.2.8 Behandlung von fehlerhaftem Input

Grundsätzlich ist für jede Eingabe eine möglichst regeltreue Analyse zu suchen. Wenn ein Satz strukturell mehrdeutig ist, ist diejenige zu wählen, die deutlich weniger Regeln verletzt.

- Dreiviertel der Unternehmen in der Internet-Industrie werden/VAFIN aufgekauft/VVPP oder/KON scheitern/VVFIN. (nicht VVINF!)

Enthält der Eingabesatz unabweisbar Fehler, die offensichtlich auf Nachlässigkeit zurückzuführen sind, so wird diejenige Struktur annotiert, die dem wahrscheinlich intendierten Satz am ähnlichsten ist. Beispielsweise soll ein falsch geschriebenes Verb als ein unbekanntes Verb angesehen werden, ein falsch geschriebenes Nomen als unbekanntes Nomen etc.

- Der Versuch, 512 Linux-Systeme zu einem Cluster zusammenzufasssen/OBJI, ist in der Nacht zum Sonntag an der Uni Paderborn erfolgreich beendet worden.

Wenn ein Wort offensichtlich fehlt oder wiederholt ist, so sind die nicht einzuordnenden Bestandteile als Fragment anzusehen.

- Zum ersten Mal ist den USA/S eine Gefängnisstrafe gegen einen Kinobesucher verhängt worden, der das Leinwandgeschehen mit seiner Videokamera aufgenommen hat.

Bei direkter Wiederholung wird dasjenige Wort in den Satz eingebaut, das seinem Regenten näher steht, das andere bleibt Fragment.

- Der Internet-by-Call-Tarif ohne Anmeldung und Grundgebühr soll von/S von/PP 6 auf 4,9 Pfennig gesenkt werden.

Bewirkt ein Schreibfehler eine Kategorieänderung, durch die die ähnlichste Struktur eine harte Bedingung verletzt, so müssen eventuell mehrere Bestandteile als Fragment angesehen werden.

- Unabhängig vom Patch-Level steht also jedes System, dass/S mit Sasser infiziert ist/S, sperrangelweit offen.

Diese Regeln gelten nur dann, wenn das Ziel explizit ist, das Verhalten des Parsers auf Realdaten zu analysieren. Wenn dagegen der Zweck des Annotierens ist, ein allgemein verwendbares Korpus deutscher Syntaxstrukturen zu schaffen, sollte natürlich zuerst der Input korrigiert und dann die offensichtliche Annotation gewählt werden.

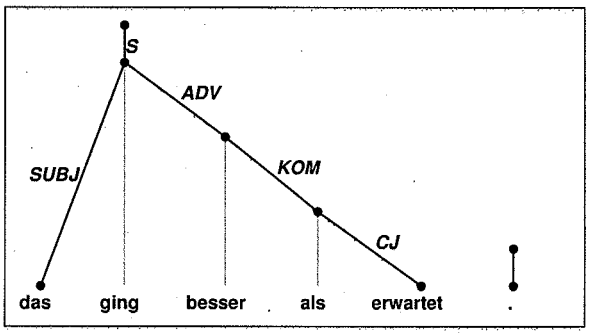
1.2.9 Ungelöste Probleme

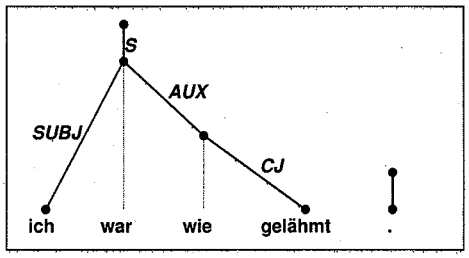
Fast alle regelmäßig vorkommenden Phänomene des Deutschen können angemessen modelliert werden. Es verbleiben jedoch Ausnahmen; einige davon sind grundsätzlich nicht im Dependenzformat zu lösen, andere wurden bislang wegen extremer Seltenheit nicht eingearbeitet.

Wenn ein einigermaßen regelhaft in akzeptablem Deutsch auftretendes Phänomen nicht richtig modelliert werden kann oder von der Grammatik hart bestraft wird, so ist das auf jeden Fall ein Fehler in der Grammatik. Alle diese Fehler sollten hier vermerkt werden, nicht im `TODO` oder sonstwo.

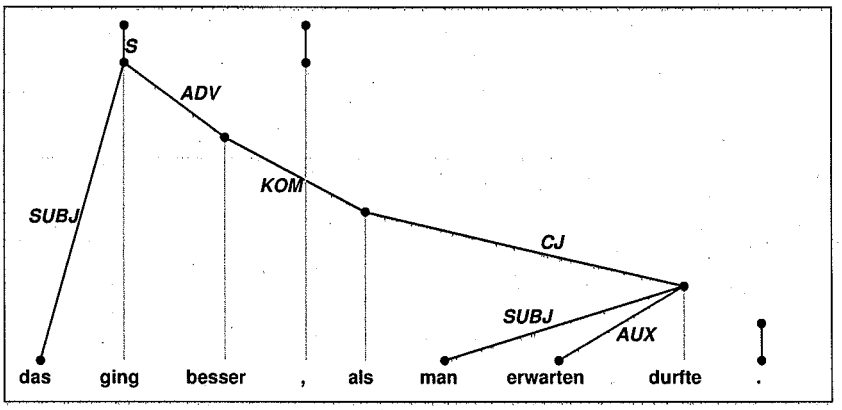
Bekannter Fehler 1: Pseudopassiv durch `KOKOM` wird nicht erkannt

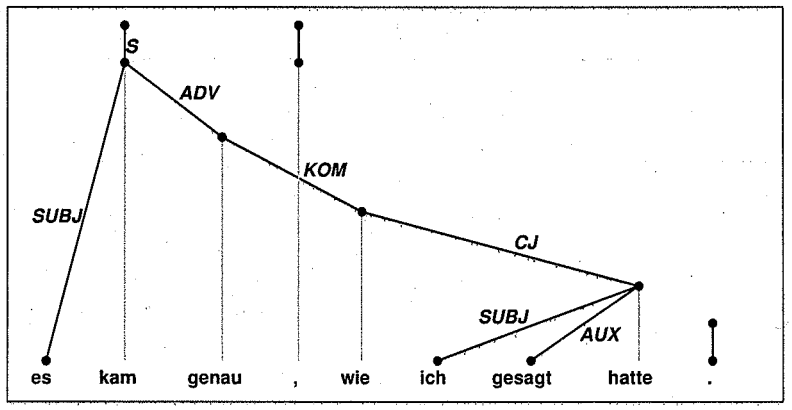
Objekte können auch dann fehlen, wenn eine Konstruktion mit 'wie' oder 'als' gebildet wird:





Hierfür wird eine Ausnahme von der Existenzbedingung gemacht, wenn ein Partizip direkt von 'wie' oder 'als' dominiert wird. Das Vergleichswort kann aber auch weiter entfernt stehen:



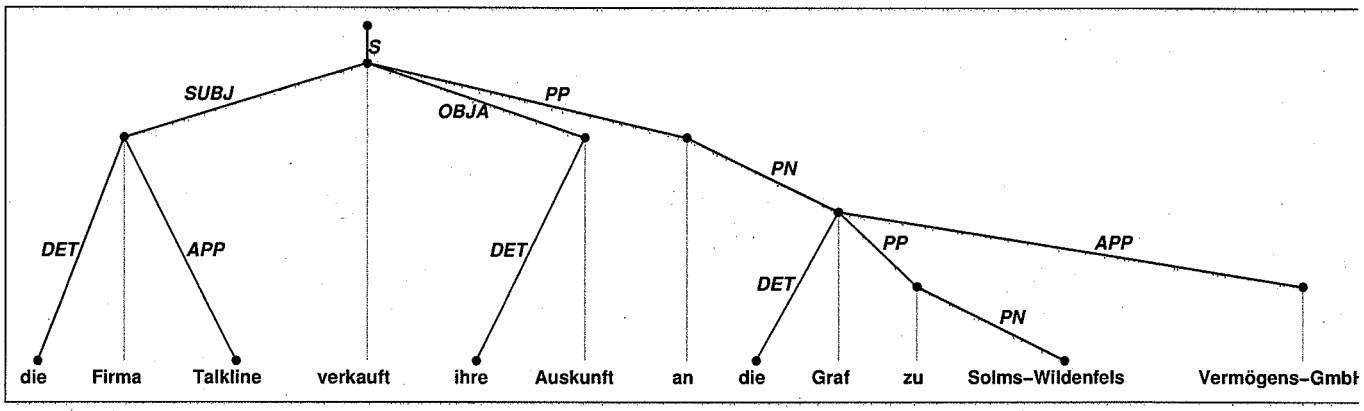


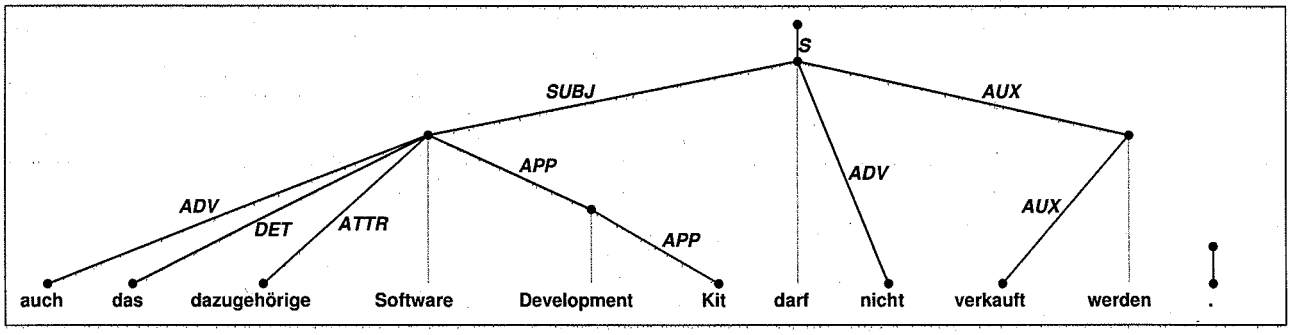
In diesen Fällen kann die Ausnahmesituation nicht erkannt werden, und die Grammatik meldet einen Fehler.

Probleme bei der Tokenisierung

Einige Probleme ich nicht als Fehler in der Grammatik an, sondern sie müßten im Tokenizer behoben werden.

Wenn Eigennamen aus normalen deutschen Worten zusammengesetzt sind, ergeben sich oft Kongruenzfehler, wenn sie Wort für Wort analysiert werden ('die Graf', 'das Software'):





Solche Namen sollten als ein einziges Lexem aufgefaßt werden, was aber eine Komponente für *named entity recognition* voraussetzt.

Alternativ könnte man den Namen anders analysieren, so daß die sinnvollere Unterordnung gewählt werden kann ('die GmbH', 'das Kit'). Dazu müßte aber die strenge Rechtsverzweigung aufgegeben werden, was die Mehrdeutigkeit sehr erhöhen würde.

Allgemein ist leider die Tendenz zu beobachten, zusammengesetzte Hauptworte (wie im Englischen) durch Leerzeichen zu bilden statt durch Bindestriche:

"Ich wollte die Charaktere *der Manson Familie* als Repräsentanten derselben Archetypen zeigen, die man in der griechischen Tragödie findet."

Derartige Komposita können nicht erkannt werden und führen dann zumindest zu Flektionsfehlern.

Wo der Bindestrich doch auftritt, wird er bisweilen durch falsche Tokenisierung wieder vernichtet:

"In 13 der 16 Stadtteilparlmente sitzen jeweils 19 Beiräte , nur in dreien (Kalbach , Harheim und Nieder-Erlenbach) kommen jeweils neun " *Vor-Ort* " *-Politiker* zu den monatlichen Sitzungen ."

Hier sollte die Zusammensetzung als ein Nomen behandelt werden.

Verschiedenes

Diverse seltene Konstruktionen sind nicht richtig modelliert. Alle folgenden Probleme sind ungelöst.

Bekannter Fehler 2: Metagrammatische Benutzung kann nicht richtig verarbeitet werden

Beispiel: 'Der Online-Versand verkauft Adolf Hitlers *Mein Kampf*.'

Hier besteht Kasus-Inkongruenz, weil das Akkusativobjekt ein Buchtitel ist, der in sich natürlich Nominativ ist. Ähnlich wie bei mehrteiligen Namen müßte der Titel als ein einziges zwar intern strukturiertes, aber nicht nach außen kongruierendes Lexem angesehen werden. Hier macht sich bemerkbar, daß die Dependenzgrammatik keine Zwischenknoten kennt.

Bekannter Fehler 3: Genitivus absolutus ist nicht vorgesehen

'Solange das Denken in einer Epoche nicht auf sein eigenes Niveau gelangt, kann man *strengen Sinnes* nicht vom Fortschritt in der Philosophie reden.'

Ein Genitivus absolutus ist ebenfalls möglich, aber nur mit ganz bestimmten Konstruktionen, die man wohl alle eigens aufzählen müßte, um sie zu erlauben.

Bekannter Fehler 4: Idiomatische Ausdrücke verletzen Syntaxregeln

Idiomatische Ausdrücken können nur dann behandelt werden, wenn sie zwar semantisch anormal sind, aber den normalen Syntaxregeln gehorchen, also Wendungen wie 'ins Kraut schießen'.

Wenn ein Idiom aber auch syntaktisch anormal ist, kann es gewöhnlich nicht vollständig analysiert werden:

'Die Chipschmieden sind gezwungen, *auf Teufel komm raus* zu verkaufen.'

'Die Chips sind *frei Hersteller* für unter 2,50 Dollar zu haben.'

Bekannter Fehler 5: Im Funktionsverbgefüge können sich Valenzrahmen ändern

Die möglichen Argumente von Verben sind jeweils im Verb selbst als Valenzrahmen kodiert. Das stellt sicher, daß etwa ein `OBJA` nur mit einem Verb auftreten darf, daß als transitiv markiert wird. Dasselbe gilt für alle anderen Nominalobjekte.[2] Dabei ist dem Phänomen Rechnung getragen, daß zusammengesetzte Verben sich anders verhalten als ihre Grundformen, selbst dann, wenn sie im Satz getrennt geschrieben werden.

Nun gibt es aber Verben, deren Valenzrahmen sich auch dann ändert, wenn sie in bestimmten festen Fügungen verwendet werden. Beispiel:

- Ich mache Spaß(OBJA).
- Ich mache dich(OBJA) glücklich(PRED).
- Ich mache dich(OBJA) zum König(PRED).
- *Ich mache zu gehen(OBJC).

Aber:

- Ich *mache Anstalten*, zu gehen(OBJC).

[2]Ein Dativobjekt darf allerdings frei auftreten (ethischer Dativ).

In diesem Fall könnte man das OBJA als Komplement zu 'Anstalten' ansehen. Aber auch andere Verben zeigen dies Verhalten:

'Ich *habe es schwer*, mich durchzusetzen.'

'Ich *bin froh*, daß das vorbei ist'.

Diese Art der Valenzänderung ist derzeit nicht repräsentierbar.

Bekannter Fehler 6: Der Objektsatz mit 'was' ersetzt das Objekt

Jedes Akkusativobjekt kann durch einen Objektsatz mit 'was' ersetzt werden:

'Was da auf dem Boden herumlag, hatten Angestellte des Hessen-Centers selber heruntergeklopft, um eben einem derartigen Steinschlag vorzubeugen.'

Die Grammatik erkennt diese Konstruktion nicht als Objekt; stattdessen ist der was-Satz ein normaler Relativsatz, und die Valenz bleibt unerfüllt.

In altertümlichem Deutsch war diese Konstruktion sogar für Relativsätze der 'das'-Reihe möglich:

'Und siehe, ich will segnen, die dich segnen.' (= die, die dich segnen)

Dieser Relativsatz kann überhaupt nicht mit dem Hauptsatz verbunden werden und bleibt fragmentarisch.

Bekannter Fehler 7: Geteilte Präpositionaladverben

In der Umgangssprache wird zum Beispiel 'damit' oft getrennt in 'da' + 'mit':

'Da kann ich mich jetzt nicht mit befassen.' 'Da hat er sich nicht zu geäußert.'

Bisweilen wird die Präposition sogar gleichzeitig gebunden und ungebunden verwendet:

'Das gehört zum modernen Leben dazu.'

Bekannter Fehler 8: Dialekt kann nicht erfaßt werden

Dialekte des Deutschen enthalten viele abweichende Formen, oft aber auch normale. Daher können sie nicht pauschal als `FM` abgeschrieben werden:

" wolle mern reilasse ? "

Kapitel 2

Die Constraintgrammatik

2.1 Constraints

Zu jedem Label FOO existieren harte unäre Constraints namens 'FOO-Definition' und 'FOO-Unterordnung', die erlaubte Regenten und Modifikatoren aufzählen. Beispielsweise erlaubt 'ADV-Unterordnung' die Unterordnung von Adverben unter alle Inhaltsworte, aber nicht Satzzeichen etc. Diese Definitions-Constraints sollten immer im strengen Sinne unär sein, d.h. keine Kontext-sensitiven Funktionen benutzen, damit sie angewendet werden können, bevor der Parser zu arbeiten beginnt.

Kanten, die zwar erlaubt, aber dispräferiert sind, müssen in eigenen Constraints bestraft werden. Wenn beispielsweise ein Infinitiv nur zusammen mit 'zu' als Objektinfinitiv erlaubt ist, sollte dennoch nicht 'has' im Definitionsconstraint verwendet werden, denn dadurch würde die Auswertung beim Netzaufbau gänzlich verhindert. Stattdessen sollte der Infinitiv erlaubt und die Forderung nach dem 'zu' in einem zweiten, Kontext-sensitiven Constraint ausgedrückt werden.

2.1.1 Namen

Alle Constraints sollten Namen in (für Linguisten) allgemeinverständlicher Sprache tragen. Der Name soll den vorliegenden Fehlerfall beschreiben, nicht den nicht erfüllten Normalzustand. Ein Anordnungsconstraint sollte also heißen 'Artikel steht rechts', nicht 'Artikel muß links stehen'.

Hier einige Muster für Constraintnamen:

`FOO-Definition`: stellt Bedingungen an Worte, die als FOO untergeordnet werden. Oft hart.

`FOO-Unterordnung`: stellt Bedingungen an Worte, denen ein FOO untergeordnet wird.

`FOO ohne Komma`: Das Phänomen FOO wird üblicherweise durch Komma markiert, das hier fehlt.

`FOO-Kasus`: Die Beziehung FOO verlangt Kasuskongruenz oder -Rektion, die hier verletzt ist.

`FOO ohne "x"`: Die Beziehung FOO muß durch das Wort "x" markiert werden, das hier fehlt.

`irreführendes FOO`: Die Abfolge von Kategorien legt eine andere Konstruktion nahe, als gewählt wurde.

`X should be Y`: Die Konstruktion kann auf zwei verschiedene sinnvolle, aber nicht bedeutungsunterscheidende Arten analysiert werden, und der Einheitlichkeit halber soll immer Y gewählt werden.

`find_FOO`: Hilfsconstraint, das nach Vorkommen des Phänomens FOO sucht.

2.1.2 Gruppen

Jedes Constraint wird ausdrücklich in einer bestimmten Gruppe ('section') deklariert. Im allgemeinen Teil entspricht diese Deklaration meistens dem entsprechenden Abschnitt der Grammatik, d.h. alle Constraints im Abschnitt 'Distanz' gehören der Gruppe 'dist' an etc. Im systematischen Teil sind Constraints aus verschiedenen Gruppen zusammen angeordnet; z.B. gibt es im Abschnitt über das Label SUBJ Constraints über Anordnung, Eindeutigkeit etc.

Folgende Gruppen gibt es derzeit:

Gruppe	Zweck
POS	Tagger-Information
agree	Übereinstimmung von Features wie in U-Grammatiken
category	Verträglichkeit von syntaktischen Kategorien
debug	Constraints mit Wert 1.0, nur zu Diagnosezwecken
dist	Constraints, die kurze Anbindung bevorzugen
exist	Existenzforderungen
help	Constraints, die nur indirekt verwendet werden
init	Strukturconstraints, deren Verletzung unsinnig ist
pref	disambiguieren, wenn keine andere Information vorliegt.
lexical	Spezialverhalten einzelner Lexeme
order	Anordnungsregeln
proj	Projektivität
punc	Stellung von Satzzeichen
root	Regeln über Wurzelanbindung
shallow	Regen anhand von Kategorieabfolgen
sort	sortale Restriktionen
uniq	Eindeutigkeit mancher Relationen
zone	Einschränkungen der Anbindung zwischen Satzzonen

Die Einordnung von Constraints in Gruppen ist oft nicht eindeutig; eine Regel kann z.B. die Anordnung zweier Worte regeln, wenn das eine in Anführungszeichen steht, und gehört dann sowohl nach 'order' als auch nach 'punc'. Leider erlaubt CDG keine doppelte Klassifizierung.

Tabellen werden zusammen mit dem Constraint angeordnet, das sie benutzt.

Viele Constraints fallen in mehrere dieser Klassen, aber CDG erlaubt keine Mehrfacheinordnung. Jedes Constraint ist deshalb der Klasse zugeteilt, der es am besten entspricht.

Die Gruppe POS

Diese Constraints nutzen Information des part-of-speech taggers, wenn vorhanden. Sie wird anhand der CDG-Funktion `pts()` gelesen.

Beispiel: der Score des Taggers wird direkt als Constraint-Score verwendet.

```
{X:SYN} : tagger : POS : [ pts(X@id) ] :
  pts(X@id) = 1.0
```

Die einzige sinnvolle Verwendung dieses Wertes ist die Forderung, daß er hoch sein soll. Statt den Wert selbst als score zu verwenden, kann man ihn auch in verschiedener Weise normalisieren.

Die Gruppe agree

Diese Constraints fordern die Übereinstimmung (Kongruenz) oder Existenz (Rektion) von Features innerhalb einer Phrase.

Beispiel: Determiner und Nomen kongruieren im Kasus.

```
{X!SYN} : DET_Kasus : agree : 0.1 :
  X.label = DET
  ->
  exists(X@case) & compatible(Kasus,X@case,X^case);
```

Das Subjekt steht im Nominativ.

```
{X!SYN} : Subjekt_Kasus : agree : 0.1 :
  X.label = SUBJ &
  exists(X@case)
  ->
  compatible(Kasus,X@case,nom);
```

Da nicht alle Lexikoneinträge voll spezifiziert sind, muß gewöhnlich `exists()` und `compatible()` verwendet werden.

Verletzungen der Kongruenz sind verhältnismäßig schwere Fehler.

Die Gruppe category

Diese Constraints fordern bestimmte Wortkategorien (feature `cat`). Beispiel: Das Subjekt ist eine NP (oder ein Subjektsatz).

```
{X!SYN} : Subjekt_Kategorie : category : 0.0 :
  X.label = SUBJ
  ->
  isa(X@,Nominal) | isa(X@,not_PP) | X@cat = ADJA;
```

Sie verwenden oft das Makro `isa()`.

Die Gruppe debug

Diese Constraints haben immer den Wert 1.0, d.h. sie beeinflussen die Analyse nicht. Sie dienen lediglich dazu, anzuschlagen, wenn eine erlaubte, aber außergewöhnliche Situation eintritt, z.B. eine Ausnahmeregelung in einem anderen Constraint, die sonst unbemerkt bliebe. Beispielsweise wird jede Projektivitätsausnahme durch ein eigenes debug-Constraint scharf bewacht.

Die Aufgabe von Debug-Constraints läßt sich in vielen Fällen auch durch `search-annotations.pl` erfüllen.

Die Gruppe dist

Diese Constraints bestrafen lange Anbindungen leicht, so daß im Zweifelsfall die kürzere gewählt wird. Sie verwenden zumeist das Makro `gradient()`.

Beispiel: von zwei PP-Anbindungen wähle die kürzere.

```
{X!SYN} : PP_Distanz : distance : gradient(100) :
  X.label = PN
  ->
  abs( distance( X@id, X^id ) ) <= 1;
```

Die Gruppe exist

Constraint, die die Existenz von Komplementen fordern. Gewöhnlich verwenden sie das Prädikat `has()`.

Beispiel: Wenn ein Verb nur mit Präfix auftritt, bestrafe dessen Fehlen.

```
{X:SYN} : 'AVZ fehlt' : exist : 0.0 :
 exists(X@avz) & X@avz = required
 ->
 has(X@id, AVZ);
```

In einigen Fällen ist die Existenz eines normalerweise notwendigen Komplementes ausdrücklich verboten. Beispiel: ein Partizip, das von 'sein' oder 'werden' abhängt, gilt als Passivkonstruktion und darf keine Akkusativobjekt tragen.

```
{X!SYN\Y!SYN} : VVPP_obl2_exist_penalty : exist : 0.0 :
 X.label = OBJA &
 isa(Y@,Partizip) &
 (Y^base = werden | Y^base = sein)
 ->
 false;
```

Die Gruppe help

Diese Constraints sind Hilfsformeln, die von anderen Constraints über has() aufgerufen werden. Sie selbst sind gewöhnlich nicht aktiv, weil sie keine allgemeingültigen Regeln beschreiben.

```
{X:SYN} : find_daß          : help : 1.0 : exists(X@Objektsatz_Konjunktion);
```

Die Gruppe init

Diese Constraints stellen Strukturbedingungen, die nicht sinnvoll verletzt werden können. Sie haben stets die Bewertung 0.0.

Beispiel: Es kann es nicht zwei Subjekte zu einem Verb geben (höchstens eine Koordination aus zwei NP).

```
{X!SYN/\Y!SYN} : 'zwei Subjekte' : init : 0.0 :
 X.label = SUBJ
 ->
 Y.label != SUBJ;
```

Die Gruppe pref

Zur Aufgabe des Parsers gehört auch die Entscheidung, welches von mehreren Homonymen im Satz anzunehmen ist. Da ein Ziel von CDG die völlige Disambiguierung ist, schreiben wir Constraints, die willkürlich einzelne Feature-Werte sehr schwach bestrafen. Dadurch wird selbst dann eine Entscheidung möglich, wenn es keinen Kontext gibt, etwa in Fragmentsätzen:

- Was, uns/acc?
- *Was, uns/dat?

Die Constraints dieser Gruppe haben Bewertungen nahe 1. Sie dienen also nur zur Disambiguierung, wenn keine andere Information vorliegt.

Die meisten Constraints dieser Gruppe betreffen nur NIL-Kanten. Das hat zwei Gründe:

1. Worte, die nicht Wurzeln sind, besitzen oft genug Kontext, daß eine Entscheidung zwischen den Homonymen möglich ist. Subjektkanten etwa unterliegen sowohl Numerus- als auch Kasusregeln.
2. In anderen Fällen wäre ein Disambiguierungsconstraint das einzige in der ganzen Grammatik, das überhaupt auf ein bestimmtes Feature zugreift. In diesem Fall ist es nicht sinnvoll, eine Unterscheidung zu treffen, denn dadurch wird die Optimierungsfunktion des Parsers, mehrere logische Varianten von Kanten zu einer Datenstruktur zusammenzufassen, außer Kraft gesetzt. Das so entstandene größere Problem ist meistens nur mit Mehraufwand zu lösen.

Beispiel: Wenn nichts anderes bekannt ist, nimm den Singular an.

```
{X:SYN} : pl : pref : 0.999 :
  exists(X@number) -> X@number!=pl;
```

Die Gruppe lexical

Diese Constraints beschreiben das Verhalten einzelner Lexeme genauer. Information dieser Art ist gewöhnlich lexikalisiert; z.B. gibt jedes Verb selbstbestimmt seinen Valenzrahmen vor, und die Grammatik regelt Verbkomplemente, indem sie auf diesen Rahmen zugreift statt auf die Wortform. In manchen Fällen ist das Verhalten aber so abweichend, daß eigene Constraints sinnvoller sind.

Beispiel: Adverben geben durch ihr Feature `modifies` an, welche Kategorien sie gerne modifizieren. Um die Fügung 'noch ein' zu modellieren, erlaubt 'noch' also unter anderem auch 'ART'. Aber der betreffende Artikel muß unbestimmt sein; diese Forderung läßt sich nicht durch `modifies` ausdrücken, also ist ein wortspezifisches Constraint nötig:

```
{X/SYN} : 'noch-Unterordnung' : lexical : 0.0 :
 X.label = ADV &
 X@word = noch &
 X^cat = ART
 ->
 X^definite = no;
```

Die Gruppe order

Diese Constraints schränken die Reihenfolge von verwandten Knoten ein.

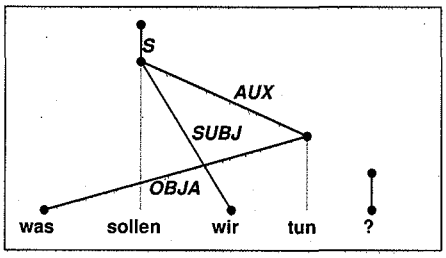
Beispiel: Der Determiner steht vor dem Bezugswort.

```
{X\SYN} : DET_order : order : 0.01 :
  X.label != DET;
```

Die Gruppe proj

Diese Constraints fordern die Projektivität von Syntaxstrukturen. Unglücklicherweise bestehen im Deutschen eine Reihe von Ausnahmen von diesem Prinzip.

Beispiel: Sinnvollerweise ordnet man das Subjekt dem finiten Verb unter, weil zwischen diesen beiden die Rektionsbeziehung besteht. Ebenso wird das Objekt dem Vollverb untergeordnet, weil dort die Valenzinformation verfügbar ist. Wenn aber das Objekt topikalisiert ist, kann es die Projektivität verletzen:



Es gibt eine Vielzahl solcher Situationen. Die meisten hängen entweder mit komplexen Verbgruppen oder mit bestimmten Worten zusammen, die die Satzstruktur beeinflussen ('und').

Alle diese Möglichkeiten erfordern Ausnahmen in der sonst unverletzlichen Bedingung. Die typischen Bedingungen zur Herstellung von Projektivität (keine interne Anbindung, keine überlappende Anbindung) werden deshalb weiter je nach der Richtung der betroffenen Kanten unterschieden und als sechs verschiedene Constraints formuliert, von denen jedes mehrere Ausnahmen hat.

Die einzelnen Ausnahmen sind jeweils mit Beispielen versehen, vergl. die Constraintdefinitionen selbst.

Als Warnung vor übergenerierenden Ausnahmen gibt es außerden debug-Constraints, die bei jeder nichtprojektiven Struktur anschlagen. Weiterhin sind bestimmte Ausnahmen weniger gern gesehen als andere; beispielsweise gilt die Auxiliar-Topikalisierung als völlig normal bzw. ein Artefakt der VP-Modellierung und wird daher nur durch ein debug-Constraint erfaßt. Dagegen ist eine nichtprojektive Koordination als markiert anzusehen und wir daher leicht bestraft:

```
{X\SYN\Y/SYN} : 'entfernte Koordination 2' : proj : 0.8 :
   Y^from < X@from
   ->
   X.label != KON & X.label != APP;
```

Die Gruppe punc

Nicht in jedem Fall stehen im Input Satzzeichen zur Verfügung; daher darf die Grammatik sie nicht in die Syntaxstruktur einbauen. Sind sie aber vorhanden, geben sie wertvolle Hinweise auf die Syntaxstruktur. Alle Constraints, die Satzzeichen untersuchen, gehören der Gruppe 'punc' an. So kann man auch Text ohne Satzzeichen untersuchen, bei leicht geringerer Performanz, indem man einfach diese Constraintgruppe abschaltet.

Diese Regel meldet etwa ein fehlendes Komma für den Nebensatz:

```
{X!SYN} : 'Komma für Nebensatz fehlt' : punc : 0.1 :
  X.label = NEB
  ->
  between(X@id,X^id,"(),-");
```

Wenn Text ohne Satzzeichen geparst werden soll, muß diese Klasse deaktiviert werden.

Nur solche Constraints, die ohne Satzzeichen nicht richtig funktionieren, gehören dieser Klasse an. Regeln, die vom *Vorhandensein* eines Satzzeichens auf bestimmte Strukturen schließen, sind bei Input ohne Satzzeichen natürlich immer trivial erfüllt. Sie können daher auch in anderen Klassen stehen, etwa `order`.

Die Gruppe root

Diese Constraints regeln die Nicht-Anbindung von Worten. Beispiel: finite Verben sind legitimerweise Satzwurzeln, aber wenn es zwei finite Verben gibt, sollte das eine ein Nebensatz zum anderen sein.

```
{X|SYN, Y|SYN} : 'mehrere Hauptsätze' : root : 0.6 :
  isa(X@,finit) &
  isa(Y@,finit)
  ->
  false;
```

Die Gruppe shallow

Diese Constraints versuchen, nur aufgrund der Abfolge von Kategorien im Satz bestimmte Unterordnungen zu verbieten. Der Gedanke dahinter ist, daß bestimmte Wortfolgen sehr stark eine bestimmte Erwartung im Hörer wecken, und eine andere als die normale Struktur stark irreführend wirkt. Sie ist daher, wenn nicht falsch, so doch ungewöhnlich und zu vermeiden.

```
{X:SYN,Y:SYN} : 'Ordinal und Monat' : shallow : 0.1 :
  X@cat = ADJA & exists(X@pattern) & X@pattern = ordinal &
  Y@cat = NN & Y@sort = month
  X@to = Y@from &
  ->
  X\Y;
```

Die Gruppe sort

Diese Constraints stellen sortale Restriktionen auf.

Beispiel: 'sein' erlaubt ein Adverb als prädikative Ergänzung, aber nur bestimmte; so ist etwa 'hier' erlaubt, aber 'heute' verboten.

```
{X!SYN} : 'falsches Adverb für "sein"' : sort : 0.0 :
 X.label = PRED &
 X^base = sein &
 X@cat = ADV
 ->
 X@subcat = local | X@base = so | X@word = umsonst |
 X@word = zuviel | X@word = anders | X@word = allein | X@word = durcheinander;
```

Die Gruppe uniq

Diese Constraints drücken Eindeutigkeit aus, zumeist unter den Labeln von nebengeordneten Worten.

```
{X!SYN/\Y!SYN} : 'doppeltes GRAD' : uniq : 0.0 :
  X.label = GRAD -> Y.label != GRAD;
```

Die Gruppe zone

Diese Constraints verbieten Anbindungen, die durch Kommas markierte Zonen in ungültiger Weise überschreiten. (Sie gehören nicht der Klasse punc an, weil sie auch für Input ohne Satzzeichen gelten.)

```
{X/SYN/\Y/SYN} : 'Objektsatz-Zone' : zone : 0.0 :
  X.label = OBJC &
  Y@from > X@from &
  between(X^id,X@id,",")
  ->
  between(X@id,Y@id,",:-");
```

2.1.3 Gewichte

Obwohl CDG ein multiplikatives Bewertungsmaß verwendet, so daß mehrfache Fehler stärker ins Gewicht fallen können als ein einzelner starker Fehler, läßt sich doch auch das Verhalten von maximierenden Fehlermaßen annähern: wählt man scores in verschiedenen Größenordnungen, dann sind auch mehrere schwache Fehler immer noch besser als ein viel stärkerer. Diese Technik wird ausgiebig angewandt. Folgende Bereiche von scores treten auf:

1.0	nur debug-Constraints
0.9999 bis 0.99	Präferenz-Constraints
0.99 bis 0.5	heuristische Regeln, die öfter verletzt werden
0.5 bis 0.1	minder schwere Fehler
0.1 bis 0.001	schwere Fehler
0.0	Wohlgeformtheitsregeln der Grammatik selbst

2.1.4 Makros für Constraintformeln

In den Constraints häufig wiederkehrende Konstruktionen sind als Makros definiert. Folgende Definitionen gibt es derzeit:

`KON_PP(L)`: wahr, wenn der Lexemknoten L eine Präposition ist, die sich wie eine Koordination verhält ('bis', 'statt').

`Partizipialadjektiv(L)`: wahr, wenn der Lexemknoten L ein Adjektiv mit dem Feature `partizipial1` oder `partizipial2` repräsentiert.

`case_specified(L)`: wahr, wenn der Lexemknoten L voll Kasus-spezifiziert ist (also nicht etwa `bot` oder `nom_acc` trägt).

`check_case(L1,L2)`: wahr, wenn die beiden Lexemknoten im Kasus kongruieren, also entweder direkt kompatibel (oder gar identisch) sind oder beide mit demselben atomaren Wert kompatibel sind (etwa `nom_acc` und `nom_dat`).

`check_gender(L1,L2)`: wahr, wenn die beiden Lexemknoten im Genus kongruieren, also entweder direkt kompatibel (oder gar identisch) sind oder beide mit demselben atomaren Wert kompatibel sind.

`edge(E, L)`: wahr, wenn das Label der Kante E in der Hierarchie `Label` dem Zweig L angehört.

`gradient(N)`: erzeugt einen Penalty-Term für ein Distanz-Constraint mit Gradient `N`.

`initial(L)`: wahr, wenn der Lexemknoten L ein Wort ist, das aufgrund der deutschen Makrostruktur zuerst stehen muß, also etwa `PWAV` oder `PRELS`, oder ein solches Wort unter sich hat (etwa `PRELAT`). Dieser Fall muß oft als Ausnahme in Anordnungsconstraints erlaubt werden.

`topicalized(E)`: wahr, wenn die Kante E ein topikalisiertes Objekt ist, d.h. aus dem Vorfeld heraus die rechte Klammer modifiziert.

`isa(L, K)`: wahr, wenn der Lexemknoten L der Kategorie K angehört.

Es gibt wesentlich mehr immer wiederkehrende Teilausdrücke in der Grammatik, die sinnvollerweise zu Makros zusammengefaßt werden sollten.

2.2 Lexikon

Die Information über mögliche und notwendige Bindungen ist zum großen Teil lexikalisiert, d.h. in den einzelnen Lexikoneinträgen festgehalten. Hier werden die vorkommenden Features mit ihrer Bedeutung erläutert.

Die Werte von lexikalischen Features können im CDG-Formalismus die Datentypen String, Zahl oder Liste haben. Zusätzlich werden einige Features verwendet, deren Bedeutung nur darin liegt, ob daß sie überhaupt definiert sind, die also wie Boolean-Werte funktionieren. Solche Features werden von der Grammatik nur durch `exists()` zugegriffen. Der Wert selbst, wenn vorhanden, lautet immer `yes`.

Folgende Features werden im Lexikon verwendet:

`Objektsatz_Konjunktion` (Boolean): Dieses Feature tragen nur die Konjunktionen 'daß' und 'ob'. Nur diese Konjunktionen können normalerweise einen Nebensatz einleiten, der als Objekt oder Subjekt verwendet wird (statt als normaler Nebensatz).

`Satzkonjunktion` (Boolean): Dieses Feature markiert Konjunktionen, die nur ganze Sätze reihen können, nicht einzelne Phrasen. So darf etwa einem beigeordneten Verb das Subjekt fehlen, wenn es von 'oder' abhängt ist, aber nicht, wenn es von 'denn' abhängt:

- Wir sehen uns den Film an, und dann gehen wir nach Hause.
- Wir sehen uns den Film an und gehen dann nach Hause.
- Ich ging nach Hause, denn ich war enttäuscht von dem Film.
- *Ich ging nach Hause, denn war enttäuscht von dem Film.

`Stoffnomen` (Boolean): Dies Feature markiert Nomen, die Stoffe o.ä. bezeichnen und daher auch im Singular ohne Artikel auftreten, z.B. 'Gold', 'Freiheit' oder 'Musik'. (Das schließt nicht aus, daß sie dennoch einen Plural bilden können: 'das Metall'/'viel Metall'/'drei Metalle').

Stoffnomen zeichnen sich dadurch aus, daß sie mit Maßnomen (vgl. Feature `set`) zusammen Mengenangaben bilden können:

- drei Becher Milch
- *drei Becher Ball
- *drei Tische Milch
- *drei Tische Ball

Das Feature `Stoffnomen` markiert also mögliche Dependenten in dieser Konstruktion

`argcat` (Liste): Spezifiziert, welche Art Worte eine Präposition als Argument akzeptiert. Z.B. nehmen die meisten Präpositionen nur NP als Argument, aber 'außerhalb' nimmt NP oder Präpositionen. Dies Feature wird mit dem Feature `cat` des Arguments verglichen.

`argprep` (Liste): Spezifiziert, welche Präpositionstypen eine Präposition als Argument akzeptiert. Beispielsweise akzeptiert 'außerhalb' eine Präposition als Argument, aber nur 'von'. Dies Feature wird gegen das Feature `prep` des Argumentes geprüft, es ist also auch 'außerhalb davon' erlaubt.

`argword` (String): Spezifiziert, welches Wort eine Präposition als Argument akzeptiert. Dies Feature wird gegen das Feature `word` des Argumentes geprüft. Beispielsweise akzeptiert die Präposition 'über' nur 'kurz' als Argument, nicht 'kurzer'.

`avz` (required/allowed/forbidden): Dieses Feature tragen finite Verben, um anzugeben, ob sie mit einem trennbaren Verbzusatz auftreten können. Beispielsweise hätte 'gibt' den Wert 'allowed', denn sowohl 'gibt' als auch 'gibt...auf' ist möglich. 'interessiert' hat den Wert 'forbidden', weil es (gemäß unserem Lexikon) keine Zusammensetzungen bildet. Den Wert 'required' hätte ein Verb, das nur als trennbares Verb auftritt.

`base` (String): Dies Feature gibt die Grundform eines flektierten Wortes an. Bei Nomen ist das der Nominativ Singular, bei Verben der Infinitiv.

`case` (nom/gen/dat/acc): Der syntaktisch ausgeprägte Kasus an einer nominalen Form. Kann unterspezifiziert sein.

`cat` (wie in Hierarchie 'Kategorien'): Die syntaktische Kategorie des Wortes. Dies ist immer eine der Kategorien des STTS.

`cat2` (wie `cat`) Einige Worte haben im Lauf der Zeit ihre syntaktische Kategorie geändert, verhalten sich aber immer noch wie Angehörige beider Klassen. In diesem Fall wird als `cat2` diejenige Kategorie angegeben, die das Wort ursprünglich hatte. Zum Beispiel ist 'anno' im Lateinischen Nomen, im Deutschen aber Adverb. Es trägt also die Attribute `cat=ADV` und `cat2=NN`.

Eine andere Klasse von mehrdeutigen Worten sind Nomen, die auch als substantivierte Adjektive angesehen werden könnten. Solche Nomen erlauben andere Modifikationen als normale Nomen:

- Der Club Behinderter und ihrer Freunde
- *Der Club Reporter und ihrer Freunde

Sie tragen daher zur Erkennung die Features `cat:NN`, aber `cat2:ADJA`.

`clitic` (Boolean): Nur gesetzt bei finitem Verb, das ein klitisches 'es' trägt und daher kein Subjekt braucht (etwa 'gibts'). Normalerweise sollten solche Formen als zwei Worte tokenisiert werden: 'gibt'+ ''s'.

`definite` (Boolean): Dieses Feature unterscheidet bestimmte von unbestimmten Artikeln, was Auswirkungen auf die erlaubten Adjektive hat. Auch einige Pronomen tragen es.

`degree` (positive/comparative/superlative): Der Steigerungsgrad von Adjektiven und einigen Pronomen.

`deverbal` (Boolean): Dies Feature markiert Nomen, die von Verben abgeleitet sind. Beispielsweise ist 'Reise nach' eine bessere Kombination als 'Tisch nach'; der Unterschied liegt zum Teil darin begründet, daß 'Tisch' atomar und 'Reise' abgeleitet ist.

`extraobjp` (Liste): Dies ist eine Liste aller Präpositionstypen, die ein Verb als `OBJP` akzeptiert, wenn es mit einem `AVZ` zusammensteht. Beispiel:

- Wir hängen den Viehdieb.
- Wir hängen den Viehdieb mit/PP Draht.
- *Wir hängen den Viehdieb mit/OBJP Draht.
- Das Ergebnisse hängen mit/OBJP Rundungsfehlern zusammen.

Folglich besitzt 'hängen' das Feature `extraobjp:mit`.

Dies Feature ist redundant, da die Verträglichkeit von Verb, Präfix und Objektpräposition auch noch durch ein ternäres `lookup()` überprüft wird. Es existiert nur, damit OBJP-Kanten schon durch ein unäres statt durch ein binäres Constraint verboten werden können, also aus Performanzgründen.

`extrasubjc` (Boolean): Dies Feature markiert Verben, die unmarkierte Subjektsätze erlauben, wenn sie mit einem bestimmten Präfix zusammenstehen. (Vgl. `extraobjp` für eine Erklärung.)

`extravalence` (String): Dies Feature gibt Kasusrahmen an, die ein Verb nur zusammen mit einem bestimmten Präfix erlaubt. (Vgl. `extraobjp` für eine Erklärung.) Die Bedeutung der einzelnen Buchstaben ist dieselbe wie im Feature `valence`.

`flexion` (weak/mixed/strong): Die Flektionsart von Adjektiven. Kann unterspezifiziert sein.

`fusion` (Boolean): Dies Feature markiert Konjunktionen, die mit ihrem Argument verschmolzen sind, etwa 'usw.'.

`gender` (fem/masc/neut): Das Genus von nominalen Formen. Kann unterspezifiziert sein.

`impersonal` (Boolean): Dies Feature markiert Vollverben, die ohne Subjekt gebraucht werden können, z.B. 'grausen'.

`invariant` (Boolean): Dies Feature markiert Adjektive, die nicht gebeugt werden, z.B. 'lila'.

`leftpenalty` (Zahl): Die Strafe, die angewendet werden soll, wenn dieses Wort als Modifikator links steht.

`likes_comparative` (Boolean): Dieses Feature tragen Adverben und Adjektive, die jede Art von Komparativ modifizieren können, sogar Attributivpronomen ('erheblich', 'viel' etc.).

`likes_positive` (Boolean): Dieses Feature tragen Adverben und Adjektive, die die Pronomen 'viele' und 'wenige' modifizieren können ('sehr', 'ganz' etc.).

`linker_Teil` (Boolean): markiert die einleitenden Teile von mehrteiligen Konjunktionen ('weder', 'je').

`loctype` (lative/essive): Unterscheidet Richtungs- von Orts-Ausdrücken.

`measurement` (Boolean): markiert Adjektive, die Gradangaben auch im Positiv zulassen ("300 Mhz schnell").

Ebenfalls markiert werden Adverben, die durch Zusammensetzung aus Nomen entstanden sind, die Gradangaben zulassen, z.b. '1000-mal', denn diese erlauben andere Modifikatoren als normale Adverben:

- rund 1000-mal
- *rund jeweils

`modifies` (Liste): Spezifiziert, welche Kategorien von Worten dieses Wort modifizieren kann, und mit welcher Bevorzugung.

`mood` (indicative/subjunctive1/subjunctive2): Der Modus eines finiten Verbs.

`need_expl` Dies Flag markiert Verben, die einen Objektsatz nur mit expletivem 'es' erlauben:

- Ich verschmähe es nicht, auch Hilfsarbeiten anzunehmen.
- *Ich verschmähe nicht, auch Hilfsarbeiten anzunehmen.

`nimmt_Objektsatz` (Boolean): Dies Flag markiert Adverben, die, wenn in einem Hauptsatz verwendet, eine freie Verbsubordination erlauben, also eine, die nicht von Valenzrahmen des Hauptverbs abhängt.

- Ich rechne damit, eine Begnadigung zu erhalten.
- *Ich rechne, eine Begnadigung zu erhalten.

Das Wort 'damit' trägt also das Feature `nimmt_Objektsatz`.

`nimmt_Subjektsatz` (Boolean): Dies Flag zeigt an, daß ein Verb einen normalen Hauptsatz als Subjekt nehmen kann. Nebensätze mit 'daß' sind immer als Subjekt erlaubt; dieses Flag markiert nur solche Verben, bei denen auch Sätze *ohne* 'daß' erlaubt sind:

- Mir fiel ein, daß ich noch kein Geschenk gekauft hatte.
- Ich hatte noch kein Geschenk gekauft, fiel mir ein.
- Daß ich kein Geschenk fand, versetzte mich in Panik.
- *Ich fand kein Geschenk, versetzte mich in Panik.

Das Verb 'einfallen' trägt also das Feature `nimmt_Subjektsatz`.

`number` (sg/pl): Der Numerus von Verb- oder Nominalformen.

`objа` (Boolean): Dies Feature gibt an, daß nur ein ganz bestimmtes Nomen als Akkusativobjekt in Frage kommt. Meistens betrifft dies die *figura etymologica*, wo Verb und Nomen denselben Stamm besitzen:

- den Heldentod sterben
- *die Schlacht sterben

Welche Paare von Verb und `OBJA` erlaubt sind, iot durch eine Tabelle in der grammatik geregelt (vgl. das Constraint 'OBJA-Wort').

`objp` (Liste): Dies Feature gibt die Präposition an, die ein Verb als Objektpräposition nehmen kann. Es kann auch eine Liste von einzelnen Strings sein, wenn mehrere Objektpräpositionen möglich sind.

`partizipial1` (Boolean): Dies Flag markiert Partizipien des Aktiv (die syntaktisch gesehen Adjektive sind).

`partizipial2` (Boolean): Dies Flag markiert deklinierte Partizipien des Passiv (die syntaktisch ebenfalls Adjektive sind).

`pattern` (String): Dies Feature zeigt an, aus welcher Regel für unbekannte Worte ein Lexikoneintrag hervorging. Es ist nicht nur informativ, sondern wird von der Grammatik benutzt, um z.B. verschiedene Arten von `TRUNC` zu unterscheiden.

`perfect` (sein/haben/bot): Dies Feature gibt das Hilfsverb an, mit dem ein `VVPP` verwendet werden muß. Einige Verben können sowohl mit 'sein' als auch mit 'haben' verwendet werden; diese tragen den unterspezifizierten Wert 'bot'.

`person` (first/second/third): Die syntaktische Person eines Nomens oder Verbs.

`phon` (String): Dies Feature gibt den abgetrennten Wortbestandteil eines TRUNC an (z.B. 'aus' für 'aus-').

`pos` (SOV/SVO): Dies Feature markiert den Satztyp (Nebensatzstellung oder Hauptsatzstellung), der bei einigen Verben direkt ablesbar ist.

`prefix` (String): Dies Feature lautet 'her' oder 'hin' bei denjenigen Adverben, die damit gebildet sind, z.B. 'herab'.

`prep` (vgl. Hierarchie 'Kasus'): Dies Feature nennt bei allen Adpositionen und Abarten davon (`APPRART`, `PROAV`, `PWAV`) die zugrundeliegende Adpositionsform. Beispielsweise tritt des Verb 'abhängen' gern mit der Präposition 'von' zusammen; die tatsächliche Wortform kann aber ebenso gut 'vom' oder 'hiervon' lauten. Dies Feature kennzeichnet alle diese Varianten als zusammengehörig. Der Einheitlichkeit halber gelten auch Vergleichsworte (`KOKOM`) als Präpositionen in diesem Sinne.

`rechter_Teil` (Boolean): markiert die rechten Teile von mehrteiligen Konjunktionen ('noch', 'desto').

`rightpenalty` (Zahl): Die Strafe, die angewendet werden soll, wenn dieses Wort als Modifikator rechts steht.

`set` (Boolean): Dies Wort markiert Nomen, die als Mengenangaben verwendet werden wie 'Becher', 'Gruppe' oder 'Stapel'.

Begründung: Gewisse Nomen verhalten sich ähnlich wie prenominale Zahlen, obwohl sie die Kategorie NN besitzen, also ein intrinsisches Genus besitzen, Artikel nehmen können etc. Sie werden also wie normale Nomen behandelt: in einem Ausdruck wie 'drei Becher Milch' ist 'Becher' der Regent und 'Milch' der Dependent (label `APP`). Das Feature `set` markiert also mögliche Regenten in dieser Konstruktion (vgl. auch das Feature `Stoffnomen`).

`sibilant` (Boolean): Dieses Feature markiert Eigennamen, die in Zischlauten enden. Das ist deshalb von Bedeutung, weil diese Namen manchmal den Nominativ als Genitiv verwenden.

`sort` (vgl. Hierarchie 'Sorten'): Rudimentäre Information über die Bedeutung eines Wortes.

`state_of_mind` (Boolean): Dies Feature markiert Verben, die eine Gemütsregung ausdrücken und daher untergeordnete Nebensätze lizensieren:

- Ich bin erstaunt, daß der Bau so groß ist.
- *Ich bin naiv, daß der Bau so groß ist.

'erstaunen' trägt also das Feature `state_of_mind`.

`stress` (stressed/unstressed/su/us/none): Dies Feature unterscheidet Varianten von Verben, die sich nur in der Betonung unterscheiden. Beispielsweise ist das 'übersetzen' mit betontem Präfix ein Synonym für 'Gewässer überqueren', während das mit unbetontem Präfix 'dolmetschen' bedeutet.

`subcat` (String): Dieses Feature unterscheidet verschiedene Typen innerhalb einer syntaktischen Kategorie. Beispielsweise sind alle Worte vom Typ `NE` genauer klassifiziert (Vorname, Nachname, Firmenname etc.), weil der Unterschied manchmal syntaktisch bedeutsam ist.

`suffix` (String): Dies Feature gibt die Endung eines attributiven Adjektivs an. Das Adjektiv 'letzter' trägt etwa das Feature `suffix:er`.

`tense` (past/present): Das Tempus eines finiten Verbs.

`valence` (String): Dies Feature gibt den Valenzrahmen von Verben (und verwandter Worte wie etwa Verbaladjektive) an in einer stark abgekürzten Form an. Die Werte haben folgende Syntax:

```
Valenz  ::=  '-'
Valenz  ::=  Slot [ '+' Slot ]
Slot    ::=  [ 'r' ] Symbol { Symbol } [ '?' ]
Symbol  ::=  'a' | 'b' | 'c' | 'd' | 'e' | 'g'
Symbol  ::=  'l' | 'k' | 'n' | 'p' | 'x'
```

Die Werte sind wie folgt zu lesen:

- Der Wert - bezeichnet ein gänzlich intransitives Verb.
- Die einzelnen auftretenden Buchstaben symbolisieren dabei die Label von Argumenten, die ein Verb nehmen kann:

Kürzel	Kantenlabel
a	OBJA
b	OBJA2
c	OBJC
d	OBJD
e	PRED
g	OBJG
l	OBJI
k	KOM
n	PN
p	OBJP
x	AUX

- Das Präfix 'r' drückt zusätzlich aus, daß statt beliebiger NP nur Reflexivpronomen als OBJA oder OBJD erlaubt sind.
- Die Reihung durch + drückt aus, daß beide Argumente zugleich auftreten können.
- Die Reihung ohne + drückt aus, daß beide Argumente alternativ auftreten können. Ein Verb mit der Valenz a+d darf also bitransitiv gebraucht werden, ein Verb mit der Valenz ad nicht.
- Es können nur zwei Argumentpositionen angegeben werden. Die Reihenfolge der Argumente bei bitransitiven Verben wird *nicht* durch diese Syntax festgelegt; a+d und d+a sind äquivalent.
- Wenn zwei alternative Argumentangaben kombiniert werden, so sind stets alle möglichen Kombinationen erlaubt. Es ist nicht möglich anzugeben 'Akkusativ oder Dativ, und außerdem Infinitiv oder finites Verb, aber wenn Dativ, dann nur Infinitiv'. Solche Bedingungen müssen als eigene Constraints formuliert werden (vgl. das Constraint 'Wahrnehmungsverb ohne Infinitiv').
- Optionalität eines Slots wird durch ? ausgedrückt. Es ist nicht möglich, Optionalität verschiedener Stärke auszudrücken.
- Das Subjekt gilt nicht als Argument; die Grammatik nimmt an, daß alle finiten Verbformen ein Subjekt erwarten, außer, sie besitzen das Feature clitic.

Eine Verbform darf genau dann ein Argument tragen, wenn dessen Symbol im Feature valence des Verbs auftritt. Es darf zwei Argumente tragen, wenn deren Symbole auf verschiedenen Seiten in diesem String auftauchen. Das Fehlen von Argumenten gilt genau dann als Fehler, wenn die betreffende Argumentstelle nicht ein ? trägt.

Es kann weitere Constraints geben, die Argumente verbieten; so hat etwa das Partizip 'geschlagen' ebenso wie alle anderen Formen von 'schlagen' die Valenz a, aber das Argument wird nicht realisiert, wenn es im Passiv eingesetzt wird. Die Valenzconstraints machen diese und andere Ausnahmen ausdrücklich.

Bei Verben, die mit abgetrenntem Präfix auftreten, gilt *nicht* das Feature valence, sondern ein Wert, der der Tabelle AVZ entnommen wird. Diese Tabelle gibt zu jedem Paar von Infinitiv und Präfix an, welche Valenz für diese Kombination anzunehmen ist; beispielsweise besitzt die Form 'arbeiten' die Valenz -, aber der Eintrag für die Kombination von 'arbeiten'

mit 'ab' ist a. Das Feature `extravalence` gibt gerade diejenigen Symbole an, die ein Verb durch Kombination mit Präfixen zusätzlich unterstützt.

`value` (Zahl): Dies Feature gibt den numerischen Wert von ausgeschriebenen Zahlen an.

2.3 Hierarchien

Der Verband 'Label' zählt alle Label der Syntax-Ebene auf und unterscheidet sie nach Modifikator und Komplement (vgl. unter 'Analyseebenen'). Diverse Zwischenknoten existieren, damit Bedingungen der Form

```
X.label = A | X.label = B | X.label = C
```

abgekürzt werden können als

```
subsumes(Label, A_B_C, X.label)
```

und mittels Makros weiter zu

```
edge(X, A_B_C)
```

Der Verband 'Features' enthält solche Knoten, die als Wert eines bestimmten Features auftreten können. Wichtige innere Knoten dieses Verbandes sind folgende:

'Flexion', 'Gender', 'Kasus', 'Number' und 'Person' dominieren die verwendeten Symbole für die typischen morphosyntaktischen Merkmale, etwa 'sg' und 'pl'. Auch verschiedene teilweise spezifizierte Werte sind möglich, etwa 'nom_acc'.

Der Knoten 'Präpositionen' dominiert all jene Symbole, die als Wert des Features 'prep' auftreten können.

Der Knoten 'Kategorien' dominiert enthält alle möglichen Werte des Features `cat`. Dies sind gerade die im Stuttgart-Tübinger Tagset definierten Tags. Darüberhinaus enthält der Verband diverse Zwischenknoten, die verschiedene gröbere Klassifikationen bezeichnen. So ist z.B. die Kategorie `VVINF` sowohl 'Vollverb' als auch 'Infinitiv', und `VMFIN` sowohl 'Modalverb' als auch 'finit'.

Der Verband 'Sorten' enthält derzeit eine Minimalklassifikation von Worten nach sehr allgemeinen Bedeutungsklassen und muß vollständig abgeändert werden, bevor er nützlich sein kann.

2.4 Beispieläußerungen

Es existiert eine Sammlung über 2000 analysierten Sätzen, die die Wirkungsweise jedes Constraints demonstrieren sollen. Jeder der Sätze, die in Kommentaren vor oder in einer Constraint-Definition stehen, ist als ein Baum im Korpus 'Beispiele' annotiert. Grundsätzlich existiert zu jedem Constraint mindestens ein Satz, der die Regel erfüllt und ein verwandter,

der gegen die Regel verstößt. (Diese Zuordnung ist allerdings nicht ganz vollständig.) Wo eine Constraintformel komplex ist oder verschiedene Fälle nacheinander behandelt, sind auch mehr Beispiele zu einem Constraint vorgesehen.

Durch Auswertung der Grammatik auf diese Korpus kann geprüft werden, ob eine vorgenommene Änderung genau die beabsichtigte Wirkung hatte. Dadurch können Regressionen (das erneute Begehen bereits beseitigter Fehler) weitgehend vermieden werden. Idealerweise sollte die Änderung eines Constraints nur auf seine eigenen Beispielsätze Auswirkungen haben; daher sollte ein Beispielsatz, ob positiv oder negativ, stets so formuliert werden, daß er möglichst vielen Regeln genügt. (Allgemeine Präferenzregeln wie 'lieber Nominativ als Akkusativ' werden natürlich sehr oft verletzt; dagegen ist nichts zu machen.) Er sollte also so kurz als möglich sein, um das Phänomen beschreiben zu können, und inhaltlich möglichst prägnant, so daß die beabsichtigte Struktur auch ohne gezeichnete Pfeile deutlich wird.

Kapitel 3

The Making of deutsch.cdg

3.1 Die Dateien und ihre Bedeutung

Man kann die Grammatik des Deutschen verwenden, indem man die Datei `deutsch.cdg` mit dem Befehl `load` lädt. Üblicherweise tut man dies durch den Aufruf

```
cdg deutsch
```

oder

```
xcdg deutsch
```

Alle Constraints, Lexikoneinträge etc. sind in dieser Datei entweder enthalten oder durch *Autoloading* verfügbar, das durch Befehle in dieser Datei eingeschaltet wird. Auch die für die Benutzung des Parsers empfohlenen Hilfsprogramme werden hier festgelegt. Beispielsweise wird dringend empfohlen, statistisches Kategorietagging zu verwenden, um unplausible Homonyme abzuweisen; daher enthält `deutsch.cdg` unter anderem die Befehle

```
#pragma set taggerCommand deutsch-tagger.pl
#pragma tagger on
```

Andere Hilfsprogramme können folgen.

Diese Datei ist also für die reine *Verwendung* der Grammatik ist ausreichend. Für die *Erweiterung* sind jedoch wesentlich mehr Ressourcen notwendig. Hier ein Überblick über die beteiligten Dateien:

Datei oder Verzeichnis	enthält
AVZ.cdg	automatisch erzeugte Valenzconstraints
Adjektiv-Templates.txt	aus Templates erzeugte Adjektive
Adjektiv-Templates.cdg	cdg-Version von Adjektiv-Templates.txt
Adjektive.txt	ausdrücklich deklarierte Adjektive
Adjektive.cdg	cdg-Version von Adjektive.txt
ExtraNachnamen.txt	zusätzliche (unbenutzte) deutsche Nachnamen
Grammatik.cdg	handgeschriebene Constraints
Hierarchien.cdg	Verbände von Attributwerten
Lexikon.cdg	Worte geschlossener Wortklassen
Namen.txt	ausdrücklich deklarierte Eigennamen
Namen.cdg	cdg-Version von Namen.txt
Nomen.txt	ausdrücklich deklarierte Nomen
Nomen.cdg	cdg-Version von Nomen.txt
TODO	Hinweise von Entwicklern für Entwickler
Verben.txt	ausdrücklich deklarierte Verben
Verben.cdg	cdg-Version von Verben.txt
base.m4	Programmeinstellungen für CDG
base.cdg	Expansion von base.m4
deutsch-lexikon.cdg	Konkatenation ausdrücklich definierter Worte
deutsch-lexikon.db	Index für deutsch-lexikon.cdg
deutsch.m4	die gesamte Grammatik
deutsch.m4	Expansion von deutsch.m4
doc.tex	diese Anleitung
extra.m4	Templates für offene Wortklassen
extra.m4	Expansion von extra.m4
make-adjectives.pl	Übersetzer für Adjektive.txt
make-names.pl	Übersetzer für Namen.txt
make-nouns.pl	Übersetzer für Nomen.txt
make-verbs.pl	Übersetzer für Verben.txt
makros.m4	Makros für Constraintformeln
postscript	die Illustrationen dieses Dokumentes

Die Datei `deutsch.cdg` wird aus verschiedenen Gründen automatisch erzeugt und zusammengesetzt. Zum einen enthalten Lexikon und Constraints Makros, die nicht Teil der CDG-Eingabesprache sind, sondern erst mit dem Präprozessor m4 expandiert werden müssen. (Da CDG automatische Expansion unterstützt, ist es auch möglich, die Datei `deutsch.m4` zu laden, aber das bedeutet, daß alle Expansionen bei jedem Laden wieder ausgeführt werden müssen.) Zum anderen können so verschiedene Varianten der Grammatik erzeugt werden; beispielsweise kann es sinnvoll sein, für ein bestimmtes Korpus nur diejenigen Worte ins Lexikon aufzunehmen, die darin tatsächlich vorkommen. In diesem Fall werden `Namen.cdg`, `Verben.cdg` etc. in entsprechend eingeschränkter Weise erzeugt und dann mit dem Rest der Grammatik zu einer spezialisierten Datei vereint. Deshalb sollten bei der Arbeit an der Grammatik in jedem Fall nur die Ausgangsdateien verändert werden, also `Verben.txt` statt `Verben.cdg`, `deutsch.m4` statt `deutsch.cdg` etc.

3.2 Adjektive.txt und Adjektive.cdg

Die Datei `Adjektive.txt` enthält alle Deklarationen für Adjektive (außer Verbaladjektive, die aus `Verben.txt` erzeugt werden).

Die Deklarationssyntax verwendet das Schlüsselwort 'adj' für die Adjektivdeklaration und 'variant' für einzelne Nebenformen (vgl. 3.5). Verschiedene Flektionsklassen gibt es bei Adjektiven nicht, lediglich einige phonetische Abweichungen werden durch Zusatzangaben deklariert.

Unregelmäßigkeiten bei Adjektiven gibt es nur für die Steigerungsformen. Wo nötig, werden die drei Grundformen ausdrücklich angegeben:

```
adj gern,lieber,liebsten
```

Adjektive, bei denen auch die attributive Form unregelmäßig ist, geben vier Formen an:

```
adj hoch,hoh,höher,höchsten
```

Alle Zusatzangaben zu Adjektiven werden als einzelne Buchstaben notiert. Folgende Möglichkeiten gibt es:

- `U`: Steigerung mit Umlaut ('alt/älter')
- `D`: Doppelte Steigerung, d.h. sowohl mit als auch ohne Umlaut ('nass/nasser/nässer')
- `A`: nur attributiv ('inner')
- `E`: Adjektive auf -el, -er, -en, die bei Deklination im Grundform das 'e' verlieren können ('teuer/teuere/teure')
- `I`: gänzlich invariante Adjektive ('lila')
- `N`: Adjektive, die immer als Nomen verwendet werden ('Gesandter')
- `O`: Adjektive, die nur im Positiv vorkommen ('riesengroß')
- `P`: nur prädikative Adjektive ('allein')
- `S`: Adjektive, die nur im Superlativ vorkommen ('viertgrößt')
- `T`: Adjektive, die als Template für viele verschiedene Adjektive dienen ('[0-9]+malig')

Adjektive zu Städtenamen, die aus mehreren Teilen bestehen, sollten als ein ADJA gelten. CDG verwendet problemlos Vollformen, die Leerzeichen enthalten; der Übersetzer erlaubt jedoch keine Leerzeichen in Deklarationen. Stattdessen kann ein Unterstrich geschrieben werden:

```
adj Bad_Nauheimer
```

Adjektive werden bisweilen groß geschrieben, um Eigennamen, Titel etc. hervorzuheben. Dies wird von der Grammatik erlaubt; Phrase wie 'der Hohe Priester' oder 'die Deutsche Bank' können also normal verarbeitet werden. Es sollte keine eigene großgeschriebene Variante deklariert werden.

Deklariert werden hingegen solche Adjektive, die gewohnheitsmäßig groß geschrieben und wie Nomen eingesetzt werden: der Beteiligte, die Geliebte, das Positive. Solche Worte werden mit dem Flag `N` deklariert.

Adjektive, die nicht mehr wie Adjektive dekliniert werden, werden dagegen als echte Nomen in `Nomen.txt` deklariert. Beispielsweise lautet der Dative von 'Diagonale' nicht mehr 'Diagonalen', sondern Diagonale; es handelt sich daher um ein normales Nomen der Flektionsklasse `f16`.

3.3 Adjektiv-Templates.txt

Diese Datei enthält Deklarationen von *Templates*, also Lexikoneinträgen, die viele gleichartige Lexeme erzeugen. Beispielsweise werden Adjektive wie 'herrenlos' oder 'abstandsmäßig' durch folgende Templates abgedeckt:

```
adj [a-zäöü].+los      T 0 pattern:los
adj [a-zäöü].+mäßig    T 0 pattern:mäßig
```

Die Deklarationssyntax ist dieselbe wie in `Adjektive.txt`.

3.4 Namen.txt und Namen.cdg

Die Datei `Namen.txt` enthält alle Deklarationen für Eigennamen (Worte der Klasse `NE`). Die Syntax der Deklarationen ist

```
Datei         ::= { Kommentar | Deklaration }
Kommentar     ::= '#' String Zeilenumbruch
Deklaration   ::= Subkategorie String { Angabe }
Subkategorie  ::= 'Firma' | 'Vorname' | 'Nachname' | 'Produkt' | 'Region'
Angabe        ::= 'fix' | 'fem' | 'masc' | 'plu'
```

Eigennamen werden sehr grob in verschiedene, Syntax-relevante Subkategorien eingeteilt. Dabei gelten folgende Richtlinien:

- Firma: Dies sind Firmen, Institutionen oder Organisationen (Dell, ETH, Nato, FDP)
- Nachname: Nachnamen von Personen (Müller)
- Produkt: Namen von spezifischen Produkten (Gameboy)
- Region: Dies sind Städte, Länder, Flüsse und andere benannte geographische Elemente (Hamburg, Nordrhein-Westfalen, Elbe)

- Vorname: Vornamen von Personen (Dennis)

Einige weitere Angaben zu Namen sind möglich:

- fem: dieser Name ist grammatisch femininum (Rita, Elbe, Lufthansa).
- fix: dieser Name kann nicht gebeugt werden (BGH), insbesondere trägt er auch im Genitiv keine Endung.
- plu: dieser Name ist immer Plural (Beatles, USA).
- masc: dieser Name ist grammatisch masculinum (Hans, Gameboy).

3.5 Nomen.txt und Nomen.cdg

Die Datei `Nomen.txt` enthält alle Deklarationen für Nomen (Worte der Klasse `NN`). Die grundlegende Syntax ist

```
Datei          ::= { Kommentar | Deklaration }
Kommentar      ::= '#' String Zeilenumbruch
Deklaration    ::= ( Nomen | Variante ) Zeilenumbruch
Variante       ::= 'variant' String String
Nomen          ::= 'noun' String Klassen { Angabe }
Klassen        ::= Klasse { '/' Klasse }
Klasse         ::= ('f' | 'm' | 'n') Zahl
Angabe         ::= 'sg' | 'pl' | 'sto' | ('obj' ':' {Objektsatztyp})
Objektsatztyp ::= 'c' | 'i' | 's'
```

Die Klasse eines Nomens gibt sowohl das Genus als auch die Flektionsweise an. Sie basiert auf der in [Nakov et al. 2002] angegebenen Einteilung und wurde noch um mehrere Klassen erweitert. Hier ein Überblick über die verwendeten Klassen:

Die genaue Definition jeder Klasse ist der Tabelle in `make-nouns.pl` zu entnehmen.

Folgende weitere Angaben können zu einer Deklaration treten:

- `sg`: *singulare tantum*, es werden keine Pluralformen erzeugt.
- `pl`: *plurale tantum*, es werden keine Singularformen erzeugt. (Angegeben werden muß aber dennoch die hypothetische Singlarform, damit die Formenerzeugung richtig funktioniert, also z.B. 'noun Ferie f16 pl', obwhl es die Form 'Ferie' nicht gibt.)
- `sto`: Stoffnomen. Diese Angabe erzeugt das Feature `Stoffnomen:yes`.
- `obj`: Objektsatz. Der Wert kan aus den Buchstaben `c`, `i` und `s` bestehen und erlaubt dann die Anbindung von `OBJC`, `OBJI` oder `S` an das Nomen.

Klasse	Eigenschaften	Beispiel	Klasse	Eigenschaften	Beispiel
m0	unverändert	Fonds	f15b	unverändert	Spezies
m1	e-Plural	Tag	f15c	lateinischer en-Plural	Basis
m1a	se-Plural	Bus	f16	n-Plural	Blume
m1b	lateinischer i-Plural	Modus	f17	en-Plural	Zahl
m2	e-Plural mit Umlaut	Bach	f18	nen-Plural	Lehrerin
m3	er-Plural mit Umlaut	Wald	f19	substantiviertes Adjektiv	Angestellte
m3a	er-Plural	Leib	n20	e-Plural	Schaf
m4	Nullplural	Deckel	n20a	e-Plural mit Umlaut	Floß
m5	Nullplural mit Umlaut	Vater	n21	er-Plural	Feld
m6	s-Plural	Gummi	n22	er-Plural mit Umlaut	Dorf
m6a	optionaler s-Plural	Pkw	n23	Nullplural	Fenster
m7	substantiviertes Adjektiv	Bekannter	n23a	Nullplural mit Umlaut	Kloster
m7a	n-Plural mit Genitiv-ns	Gedanke	n24	s-Plural	Auto
m7b	n-Plural ohne Genitiv-s	Riese	n25	en-Plural	Bett
m7c	nur für 'Herr'	Herr	n25a	nur für 'Herz'	Herz
m8	en-Plural	Mensch	n26	substantiviertes Adjektiv	Junge
m8a	en-Plural mit Genitiv-s	Prozessor	n27	se-Plural	Begräbnis
m9	en-Plural mit Genitiv-es	Staat	n28	lateinischer en-Plural	Datum
m9a	ten-Plural	Bau	n28a	lateinischer en-Plural	Drama
m10	n-Plural mit Genitiv-s	Konsul	n28b	lateinischer en-Plural	Risiko
m11	lateinischer en-Plural	Organismus	n28c	griechischer en-Plural	Epos
m11a	italienischer i-Plural	Maestro	n28d	griechischer en-Plural	Stadion
m11b	lateinischer ces-Plural	Index	n28e	lateinischer en-Plural	Virus
m11c	griechischer en-Plural	Mythos	n28f	italienischer i-Plural	Porto
f12	e-Plural	Drangsal	n29	lateinischer a-Plural	Maximum
f13	se-Plural	Kenntnis	n29a	griechischer a-Plural	Lexikon
f14	e-Plural mit Umlaut	Nacht	n30	n-Plural	Auge
f14a	Nullplural mit Umlaut	Mutter	n31	lateinischer ien-Plural	Privileg
f15	s-Plural	Kamera	n32	unverändert	Internet
f15a	lateinischer en-Plural	Firma			

Tabelle 3.1: Verwendete Deklinationsklassen für Nomen.

Die bestehenden Einträge sind alphabetisch sortiert in verschiedene Gruppen eingeteilt (Nomen mit Objektsätzen, Nomen mit Zahlbedeutung, etc.). Es empfiehlt sich, diese Anordnung beizubehalten, um Duplikate leichter ausschließen zu können. (Anderenfalls werden sie vom Übersetzer gemeldet.)

Angesichts der Tatsache, daß deutsche Nomen durch Komposition beliebig erweitert werden können und überdies täglich neue Begriffe den Status von Gemeinnomen erlangen, kann die Liste von Nomen niemals vollständig sein. Hier Richtlinien darüber, wann ein Eintrag hinzugefügt werden sollte:

- Wenn ein Nomen in einem verarbeiteten Korpus vorkommt und nicht speziell dort erfunden wurde, ist es offenbar relevant und sollte behandelt werden.
- Wenn ein Nomen durch Zusammensetzung aus einem bekannten Nomen hervorgeht, kann es normalerweise fortgelassen werden. Beispielsweise gibt es einen Eintrag für 'Minister'; daher brauchen 'Verteidigungsminister', 'Landwirtschaftsminister', 'Energieminister' etc. nicht eingetragen zu werden. Hiervon bestehen jedoch mehrere Ausnahmen. Ein unbekanntes Nomen sollte immer dann eingetragen werden, wenn es sich merklich von seinem Simplex unterscheidet. Das gilt zum Beispiel in folgenden Fällen:
  - Viele Nomen werden mit dem Suffix '-schaft' gebildet. Diese sind feminin, also nicht aus dem Simplex 'Schaft' gebildet. Sie müssen also alle eingetragen werden.
  - Die automatische Dekomposition von CDG zieht nur das längste mögliche Suffix in Betracht. Wenn dieses nicht das Stammnomen ist, sollte das unbekannte Nomen dennoch eingetragen werden. Dies gilt insbesondere, wenn dadurch ein falscher Kasus oder Genus entsteht. Beispielsweise ist 'Partnerunternehmen' unbekannt und 'Unternehmen' bekannt. Der Dekompositionsalgorithmus nimmt aber an, daß 'Partnerunternehmen' ein Kompositum des bekannten Nomens 'Runternehmen' ist. Um das zu vermeiden, sollte 'Partnerunternehmen' als eigenes Nomen aufgenommen werden.
  - Die automatische Dekomposition behandelt nur Komposita, in denen Präfix und Suffix je mindestens vier Buchstaben lang sind, weil Hypothesen über kürzere Stammformen oft falsch sind. Beispielsweise gibt es sehr viele unbekannte Nomen auf '-ion'; diese werden *nicht* als Kompositum von 'Ion' angesehen, was auch nicht korrekt wäre. Alle diese müssen also eigens eingetragen werden.
  - Manche Komposita haben sich semantisch weit von ihrem Simplex entfernt. Beispielsweise verwendet das Kompositum 'Artilleriefeuer' einen bestimmten Sinn von 'Feuer' ('Beschuß'), der sich wesentlich vom häufigeren Sinn ('Verbrennen') unterscheidet. Dieser Unterschied kann eventuell syntaktisch bedeutsam sein; daher sollte ein eigener Eintrag erstellt werden. Dagegen ist 'Apfelbaum' eine gänzlich transparente Einengung des Begriffs 'Baum' und kann, aber muß nicht eigens aufgeführt werden.

3.6 Verben.txt und Verben.cdg

Die Menge der möglichen Verben des Deutschen ist nicht exakt abzugrenzen. Sowohl zusammengesetzte Verben (insbesondere mit Präfixen wie 'mit-', 'wieder-', 'zurück-' etc.) als auch

neue Verbstämme können zu Lücken im Lexikon führen. Durch Templates werden Defaulteinträge für unbekannte Verben zur Verfügung gestellt; so kann z.B. jede unbekannte Form auf '-en' als Infinitive eines unbekannten Verbs interpretiert werden, was oft für eine erfolgreiche Verarbeitung ausreicht. Beispielsweise tragen alle unbekannten Verben in Ermangelung genauerer Daten die sehr verbreitete Valenz 'ac?', die die häufigsten Satzstrukturen erlaubt. Dagegen kann etwa ein Verb mit einer Dativvalenz nur dann richtig verarbeitet werden, wenn es ausdrücklich im Lexikon genannt ist. Deshalb ist es im allgemeinen besser, eine erkannte Lücke im Lexikon durch einen neuen Eintrag zu schließen.

Alle Verben — regelmäßige wie unregelmäßige, Vollverben wie Hilfsverben — werden aus der Deklarationsliste `Verben.txt` automatisch erzeugt. Diese Liste enthält folgende Definitionen:

1. Paradigmendeklarationen: Diese geben z.B. an, daß das Imperfekt von 'stieben' 'stob' lautet.
2. Verbdeklarationen: Diese definieren ein einfaches oder zusammengesetztes Verb und alle seine Eigenschaften wie etwa Betonung oder Valenz.
3. Variantendeklarationen: Diese führen zusätzliche Formen ein, die derart unregelmäßig sind, daß sie durch den normalen Algorithmus nicht erzeugt werden können.

3.6.1 Paradigmendeklarationen

Jede unregelmäßige Verbreihe wird durch das Schlüsselwort `paradigm` deklariert. Das Paradigma gibt so viele Formen an, als nötig sind, um alle auftretenden Formen zu errechnen. Da es mehr und weniger unregelmäßige Verben gibt, sind dazu vier bis sechs unterschiedliche Formen notwendig, und zwar entweder

1. Infinitiv, 3.Sg.Imperfekt, 1.Sg.Konjunktiv 2, PPP:
   `paradigm denken,dachte,dächte,gedacht`
2. Infinitiv, 3.Sg.Präsens, 3.Sg.Imperfekt, 1.Sg.Konjunktiv 2, PPP:
   `paradigm geben,gibt,gab,gäbe,gegeben`
3. Infinitiv, 1.Sg.Präsens, 3.Sg.Präsens, 3.Sg.Imperfekt, 1.Sg.Konjunktiv 2, PPP:
   `paradigm wissen,weiß,weiß,wußte,wüßte,gewußt`

Das extrem unregelmäßige Paradigma für 'sein' ist in den Übersetzer fest eingebaut und wird *nicht* deklariert.

3.6.2 Verbdeklarationen

Die Verbdeklaration selbst hat die folgende Form:

```
VerbDeclaration ::=  Regularity 'verb' Base Valence Flags
Regularity      ::=  regular | semiregular | ''
Base            ::=  Infinitiv
Base            ::=  Prefix'>'Infinitiv
```

```
Base            ::=  Prefix'<'Infinitiv
Base            ::=  Prefix1'>'Prefix2'<'Infinitiv
Base            ::=  Prefix1'<'Prefix2'>'Infinitiv
Valence         ::=  '-'
Valence         ::=  Slot [ '+' Slot ]
Slot            ::=  [ 'r' ] Symbol { Symbol } [ '?' ]
Symbol          ::=  'a' | 'b' | 'c' | 'd' | 'e' | 'g'
Symbol          ::=  'l' | 'k' | 'n' | 'p' | 'x'
Flags           ::=  { Flag }
```

Grundsätzlich wird zwischen regelmäßigen und unregelmäßigen Verben unterschieden. Verben, die ohne 'Regularity' deklariert sind, gelten als unregelmäßig, wenn es eine `paradigm`-Deklaration für ihren Infinitiv gibt, sonst als regelmäßig. Wenn ein Verb regelmäßig ist, obwohl es ein gleichlautendes Paradigma deklariert wurde, ist `regular` anzugeben; wenn sowohl regelmäßig als auch unregelmäßig flektiert wird, muß `semiregular` angegeben werden. Beispiel:

```
paradigm bringen,brachte,brächte,gebracht
paradigm wenden,wandte,wändte,gewandt

# Normales regelmäßiges Verb: Imperfect = 'tanzte'
verb tanzen

# Normales unregelmäßiges Verb: Imperfect = 'brachte'
verb bringen

# Teilweise regelmäßiges Verb: Imperfekt = 'wendete'/'wandte'
semiregular verb wenden

# Nur regelmäßiges Verb: Imperfekt = 'entwendete' (nicht *'entwandte')
regular verb ent>wenden
```

Bei zusammengesetzten Verben muß angegeben werden, welche Betonung gilt, da dies Auswirkung auf den Satzbau hat: betonte Präfixe werden im Hauptsatz finiten Formen ans Satzende gestellt, unbetonte nicht. Dazu werden die Zeichen < und > in den Infinitiv eingefügt. Mnemonik: Die Spitze zeigt jeweils auf die betonte Silbe.

Es gibt fünf Möglichkeiten:

Notation	Betonung
fahren	vorletzte Silbe betont
um<fahren	'um' betont
um>fahren	'fahren' betont
über>vor<teilen	'vor' betont
weiter<ver>fahren	'weiter' und 'fahren' betont

Wenn verschiedene Verben sich *nur* durch die Betonung unterscheiden, wie eben 'umfahren', so müssen zwei Deklarationen geschrieben werden.

Bei langen zusammengesetzten Verben kann unklar sein, wie die Präfixe sich zueinander verhalten. In diesem Fall muß das Partizip Perfekt Passiv geprüft werden. Wenn das PPP mit 'ge-' gebildet wird, verhalten sich auch mehrere Präfixe wie ein einzelnes, betontes Präfix. Z.B. lautet das PPP von 'umherfahren' 'umhergefahren', also ist als Deklaration 'umher<fahren' anzugeben. Ebenso bildet 'wiederaufladen' die Form 'wiederaufgeladen', also lautet die Deklaration 'wiederauf<laden'. Dagegen bildet 'übervorteilen' die Form 'übervorteilt'; es handelt sich also um zwei Präfixe, von denen das zweite betont ist, und die Deklaration muß 'über>vor<teilen' lauten.

Die Valenz eines Verbs ist ein gültiger Wert für das Feature `valence` (vgl. 2.2).

Zusätzliche Eigenschaften des Verbs können hinter der Valenz angegeben werden. Die Eigenschaft 'Steigerung' bewirkt, daß von den Verbaladjektiven auch Komparativ- und Superlativformen erzeugt werden. Beispielsweise ist das Verbaladjektiv 'spannend' auch in der Steigerungsform 'spannender' gebräuchlich, daher muß das Verb 'spannen' mit der Eigenschaft 'Steigerung' deklariert werden. Dagegen läßt sich das Verbaladjektiv 'laufend' nicht steigern, also wird 'laufen' normal deklariert.

Alle anderen Eigenschaften werden uneverändert in die entstehenden Lexikoneinträge übernommen, müssen also gültige Attribut-Wert-Paare sein. Die häufigste zusätzliche Eigenschaft ist etwa `perfect=sein` für Bewegungsverben.

3.6.3 Varianten

Einige Stämme haben zusätzliche Nebenformen, die nicht der normalen Flektionsreihe entstammen. Zum Beispiel ist eine veraltete Variante von 'wurde' die Form 'ward'. Dies wird durch das Schlüsselwort `variant` deklariert:

```
variant wurde ward
```

Varianten, die sich nur durch die Rückung ß/ss unterscheiden, werden automatisch erzeugt und müssen *nicht* als Nebenformen deklariert werden.

3.6.4 Klitische Verben

Bisweilen werden Verben mit dem Pronomen 'es' zusammengezogen und etwa statt 'soll es' nur 'solls' geschrieben. Diese *klitische* Fügung müßte korrekterweise beim Tokenizing in zwei Wortformen getrennt werden, damit die richtige Dependenzstruktur gefunden werden kann; bei der Schreibweise 'soll's' geschieht dies auch bereits.

Die unmarkierte Variante kann dagegen nicht automatisch erkannt werden und wird gewöhnlich als nur ein Wort angesehen. Um solche Sätze dennoch parsen zu können, wird eine Form wie 'solls' als eigene Form ins Lexikon eingetragen, wenn sie auftritt. Diese Formen werden nicht automatisch erzeugt, sondern manuell von Verben.cdg nach Lexikon.cdg kopiert und um das Attribut 'clitic:yes' erweitert.

Dieser Mechanismus ist in mehrfacher Hinsicht unvollkommen:

- Das Attribut 'clitic' legitimiert eine Ausnahme im Constraint 'Subjekt fehlt', weil dies die häufigste Funktion eines 'es' nach einem Verb ist. Ein klitisches 'es' kann jedoch auch ein Objekt, PP-Kernel etc. sein; diese Funktion wird nicht antizipiert und führt zu irreführenden Konflikten, wenn sie auftritt.

- Nur diejenigen klitischen Formen, die in bisher untersuchten Sätzen auftreten, werden erfaßt.
- Die genaue Form der Einträge wird nicht automatisch auf den neuesten Stand gebracht, wenn sich Namen von Attributen, Wertemengen o.ä. ändern, weil sie nicht von `make-verbs.pl` erzeugt werden.

Für eine bessere Lösung müßte der Tokenizer sicher erkennen können, wann ein Wort ein klitisches Verb ist und in diesem Fall aus einem Wort zwei machen.

3.7 AVZ.cdg

Diese Datei wird automatisch aus Verben.txt erzeugt. Sie enthält folgende Grammatikelemente:

- Die Hierarchie `Valenzen` enthält alle tatsächlich vorkommenden Werte des Features `valence` als Blätter und definiert hilfreiche innere Knoten, die Constraints benutzen können, um lange Disjunktionen zu vermeiden. Beispielsweise subsumiert der Knoten `transitiv` genau jene Werte, deren erste Argumentposition kein `?` trägt, und der Wert `a_erlaubt` subsumiert genau jene Werte, die ein `a` enthalten.
- Die Hierarchien `Valenz_1` und `Valenz_2` projizieren die möglichen Werte jeweils auf die möglichen Label für die erste oder zweite Argumentposition. Beispielsweise gelten folgende Subsumptionen:

Hierarchie	X	Y	subsumes(H, X, Y)?
`Valenz_1`	`a+d`	`OBJA`	ja
`Valenz_1`	`a+d`	`OBJD`	nein
`Valenz_2`	`a+d`	`OBJA`	nein
`Valenz_2`	`a+d`	`OBJD`	ja

- Die Tabelle AVZ bildet Paare von Infinitiv und Präfix auf kombinierte Valenzwerte ab (siehe oben).
- Vier Constraints zu jedem Argumentlabel mit der folgenden Bestimmung (am Beispiel von `OBJA`):
  - Das Constraint 'OBJA nicht erlaubt' verbietet `OBJA` an Verben, die kein `a` im Feature `valence` tragen und kein Präfix erlauben.
  - Das Constraint 'OBJA mit AVZ nicht erlaubt' verbietet `OBJA` an Kombinationen von Verb und Präfix, für die kein `a` definiert ist.
  - Das Constraint 'OBJA ohne AVZ nicht erlaubt' verbietet `OBJA` an Verben, für die nur in Verbindung mit einem Präfix ein `a` definiert ist, die aber kein Präfix tragen.
  - Das Constraint 'OBJA weder mit noch ohne AVZ erlaubt' schließlich verbietet `OBJA` an Verben, die zwar ein Präfix erlauben, aber in keinem Fall dadurch ein `a` erwerben können.

Die Features `extravalence` und `avz` dienen lediglich dazu, das erste und vierte Constraint dieser Art als echtes (nicht kontextsensitives) unäres Constraint formulieren zu können.

- Die Lexikoneinträge für Verbpräfize in der abgetrennten Form; es existiert also zum Beispiel ein Eintrag für 'an' (und 'an-'), wenn mindestens ein Verb mit diesem Präfix definiert ist.

3.8 `make-verbs.pl` und verwandte Programme

Die vier Programme `make-adjectives.pl`, `make-names.pl`, `make-nouns.pl` und `make-verbs.pl` dienen dazu, aus abgekürzten Deklarationen vollständige Lexikoneinträge in CDG-Syntax zu erzeugen. Folgende Eigenschaften sind ihnen allen gemein:

- Sie arbeiten als *Filter*, lesen also STDIN und schreiben auf STDOUT. (Allerdings erzeugt `make-verbs.pl` außerdem eine Datei mit dem Namen `AVZ.cdg`.)
- Sie lesen Deklarationen der Art 'noun Tag' und erzeugen vollständige CDG-Lexikoneinträge wie etwa '`Tag := [cat:NN, number:sg, case:nom, gender:masc, person:third ]`'
- Nach der eigentlichen Deklaration können weitere typspezifische Angaben folgen, die in bestimmte Kombinationen von CDG-Features umgewandelt werden. Beispielsweise versteht der Nomen-Erzeuger die Angabe 'obj:c' und wandelt sie in das Feature `valence:'c?'` um.
- Außer diesen Angaben können auch Attribut-Wert-Paare direkt angegeben werden, die dann unverändert in den Output übernommen werden. Zum Beispiel trägt das Nomen 'Ahnung' nicht nur die Angabe 'obj:c', sondern auch 'deverbal:yes'.
- Der Output kann eingeschränkt werden auf diejenigen Formen, die für ein bestimmtes Korpus tatsächlich gebraucht werden. Dazu muß mit der Option '-c' eine Datei mit CDG-Lattices angegeben werden.
- Die Option '-v' kann benutzt werden, um die Formen von Nomen und ADjektiven mehr oder weniger stark zusammenzufassen. Die möglichen Werte sind 'min', 'med' und 'max'.

Weitere Eigenschaften der einzelnen Programme sind in den Abschnitten über die die Dateien `Adjektive.txt`, `Namen.txt`, `Nomen.txt` und `Verben.txt` beschrieben.

Literaturverzeichnis

[Nakov et al. 2002] Preslav Nakov, Galia Angelova und Walther von Hahn. 2002. Automatic Recognition and Morphological Classification of Unknown German Nouns. Bericht FBI-HH-B-243/02, Universität Hamburg.