

Institutionen för lingvistik  
Datorlingvistik  
Stockholms universitet

# Annoteringsvägledning SWE-TIGER

TIGER-annotering av svenska i Annotate

Yvonne Samuelsson

Version 1.1  
(Uppdaterad Juni 2007)

# Innehåll

<b>1</b>	<b>Introduktion</b>	<b>1</b>
1.1	Bakgrund . . . . .	1
1.2	Inför annoteringen . . . . .	1
1.3	Om Annotate . . . . .	2
1.3.1	Trädstrukturen i Annotate . . . . .	2
1.3.2	Annotate-fönstret . . . . .	3
1.3.3	Annotering steg för steg . . . . .	4
<b>2</b>	<b>Noderna</b>	<b>6</b>
2.1	Nominalfraser . . . . .	6
2.2	Adjektivfraser . . . . .	10
2.3	Adverbialfraser . . . . .	12
2.4	Namn . . . . .	12
2.5	Platshållarfraser . . . . .	13
2.6	Prepositionsfraser . . . . .	14
2.7	Satser och verbfraser . . . . .	16
2.8	Koordinerade fraser . . . . .	19
	<b>Referenser</b>	<b>21</b>
<b>A</b>	<b>SUC ordklasstaggar</b>	<b>22</b>

# 1 Introduktion

Detta är den första versionen av annoteringsvägledningen för TIGER-annotering av svenska. Förhoppningen är att vägledningen kommer att utökas och förbättras med tiden. Har du kommentarer, frågor eller rättelser är du välkommen att höra av dig mailledes, till yvonne.samuelsson@ling.su.se eller volk@ling.su.se.

## 1.1 Bakgrund

Annotate<sup>1</sup> är ett program för semi-automatisk annotering av korpusar, d v s en samling språkliga data i textform. Programmet skapades vid universitetet i Saarbrücken i Tyskland. Eftersom det ursprungligen har använts för tyska (det har en inbyggd tysk tagger och en parser tränad på tyska nyhetstexter) är större delen av dokumentationen samt annoteringsmanualen på och för tyska. Denna vägledning innehåller därför en kort introduktion till användningen av Annotate på svenska. Utöver detta har annoteringen anpassats till svenska konstruktioner och problem.

När jag började läsa datorlingvistik hade jag inte annoterat mer än några enstaka meningar för hand. Eftersom både c- och d-uppsatsen handlade om (parallella) trädbanker fick jag dock snabbt tillbringa många timmar med att semi-automatiskt annotera träd i Annotate. Att annotera är egentligen inte särskilt svårt, men i början tar det lite tid. Förutom att lära sig programmet och att tänka grammatiskt dyker det alltid upp konstruktioner som man är osäker på hur man ska hantera. Vi kommer nu att titta på hur programmet fungerar, hur en typisk fras ser ut samt hur några problematiska fraser bör hanteras. Det är dock du som annoterare som fattar det slutgiltiga beslutet om hur din trädbank ska se ut, ofta beroende på vad du sedan vill använda träden till. Det viktiga är att man försöker vara konsekvent.

Så, sätt igång. Övning ger färdighet!

## 1.2 Inför annoteringen

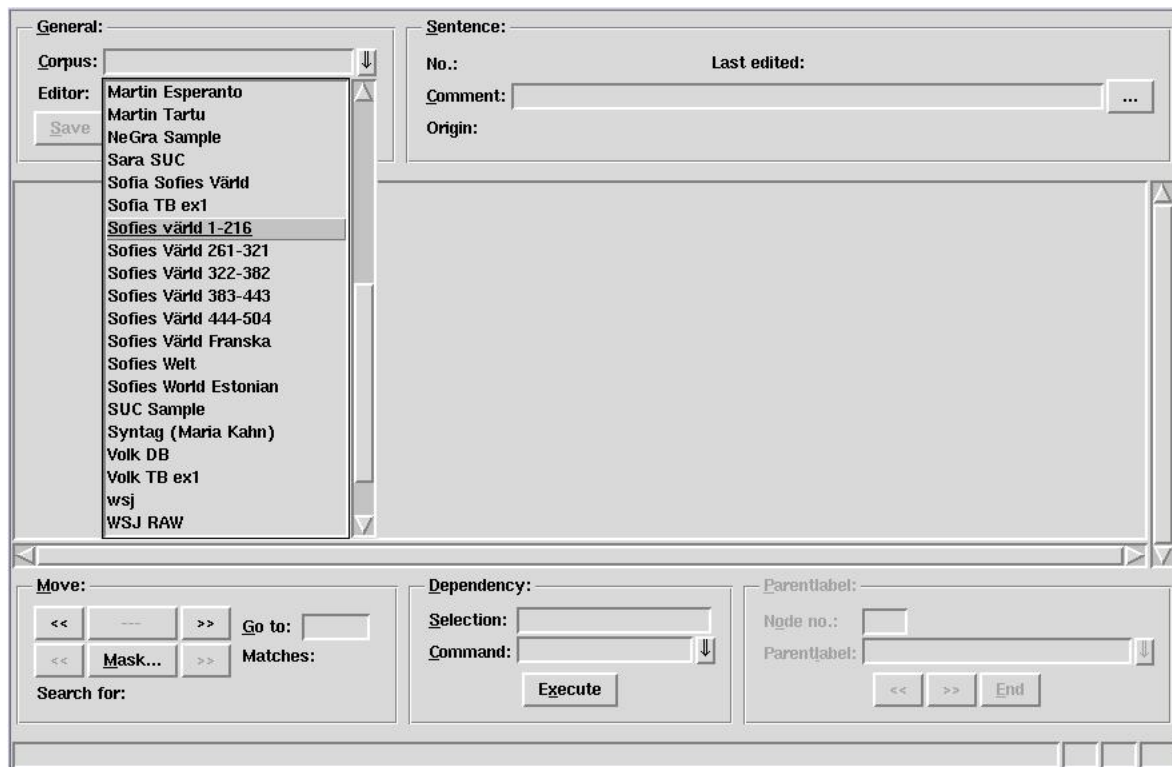
För att man ska kunna arbeta med meningarna måste de först laddas in i Annotate. De tokeniserade meningarna läggs i en databas, som hanteras av den datoransvariga på institutionen.

Innan meningarna kan parsas måste de taggas. Ursprungligen taggade vi svenska meningar innan korpusen matades in i Annotate, eftersom taggern som är integrerad i Annotate inte var tränad på svensk text. Dessutom fanns det inte någon parser tränad på svenska träd, som gav träd i enlighet med TIGER-formatet. För att lösa detta problem konverterade vi de svenska SUC-taggar till tyska STTS-taggar, så att den statistiska parser tränad på tyska texter, som medföljer programmet, tror att den arbetar med tyska meningar. (Mer information om detta kan hittas bl a i [Samuelsson 2004, Volk and Samuelsson 2004].) Det visade sig fungera bra, även om det inte är den ultimata lösningen. Denna vägledning förutsätter att parsningen görs på de svenska SUC-taggar.

Oavsett om taggningen hanteras innan meningarna matas in i Annotate, eller om de taggas i Annotate, så kan man behöva ändra vissa taggar manuellt i efterhand. En auto-

---

<sup>1</sup>Mer information finns på <http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus/annotate.html>.



Figur 1: Annotates fönster, med sex fält. Under *Corpus* i fältet *General* väljer man en korpus att arbeta med.

matisk tagger gör alltid några procent fel. En lista över SUC-taggar med beskrivning finns därför i appendix.

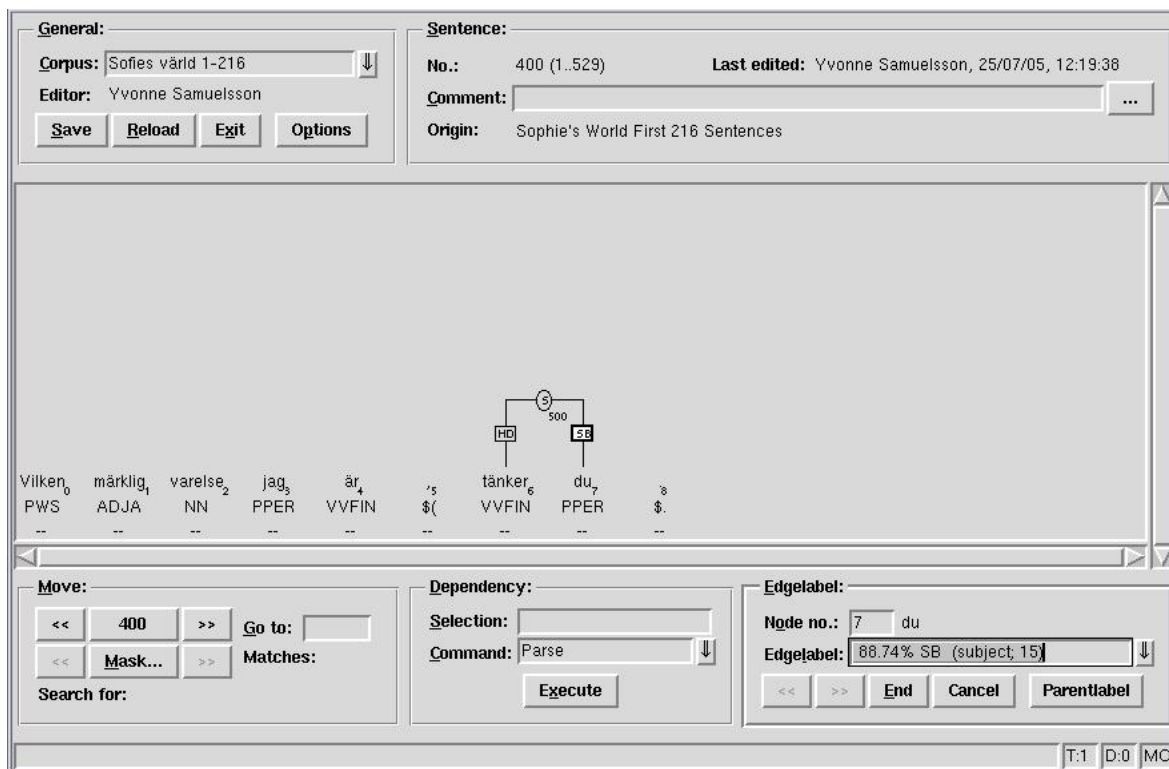
## 1.3 Om Annotate

### 1.3.1 Trädstrukturen i Annotate

Träden i en trädbank kan se ut på många olika sätt. Vi har valt att följa riktlinjerna för TIGER-annotering. Det ger en tämligen flat frasstruktur. Noderna (runda) i våra träd är fraserna och kanterna (linjerna med de fyrkantiga kantetiketterna) definierar den syntaktiska funktionen hos de olika beståndsdelarna inom frasen. Att strukturen är flat innebär att man låter bli att markera flera nodtyper som är uppenbara. Därför markeras inte finit VP, NP (nominalfras) inuti PP (prepositionsfras) eller unära noder. Unära noder definierar vi här som noder som inte förgrenas och som alltså bara innehåller ett ord eller en fras.

Det finns två stora fördelar med den flata trädstrukturen: Annoteraren måste fatta färre beslut och har en bättre översikt över träden. Nackdelen är förstås att träden inte är helt kompletta ur en lingvistisk synvinkel. Om de utelämnade fraserna å andra sidan är helt otvetydiga kan man enkelt infoga dem automatiskt i efterhand (se t ex [Samuelsson 2004, Samuelsson and Volk 2004]).

Eftersom TIGER främst är ett annoteringsformat för tyska finns det noder och kanter



Figur 2: Parsern är 88,74% säker på att *du* här är subjekt. Förslaget bekräftas eller ändras under *Edglabel*.

som inte är intressanta för svenska. Dessutom utvecklas formatet ännu, vilket innebär att det finns flera olika versioner. Jag kommer att påpeka de fall där vår annotering avviker från TIGER-formatet på något sätt.

### 1.3.2 Annotate-fönstret

Det är lätt att arbeta med Annotate och det bästa sättet att lära sig är helt enkelt att testa de olika funktionerna, "learning by doing". Här följer dock en kort introduktion till de vanligaste funktionerna.

När Annotate har startats ser man ett fönster indelat i sex olika fält. Det första fältet är *General*. Det är här, under *Corpus* man väljer en korpus att arbeta med (se figur 1). Om den korpus man väljer har arbetats med tidigare hoppar programmet automatiskt till senast uppdaterade mening vid start. I nedre delen av detta fält finns några knappar, av vilka *Save* är den viktigaste, även om Annotate automatiskt sparar eventuella ändringar när man byter till en annan mening eller går ur programmet.

Bredvid *General* finns *Sentence*, som innehåller meningens nummer (i korpusen), information om när denna mening senast bearbetades samt en kommentarsrad. Denna är bra att använda t ex om något är oklart i meningen eller meningen innehåller något man vill uppmärksamma på.

Under dessa två fält ligger trädfältet. Det är här man ser meningen och trädstrukturen. Längst ned i fönstret ligger så tre fält. I fältet *Move* kan man söka meningar på fyra sätt.

Man kan klicka fram nästa eller föregående mening med pilarna, eller hoppa direkt till en viss mening genom att skriva in dess nummer i *Go to*-fältet. Om man skriver + eller - framför siffran hoppar man istället så många meningar framåt eller bakåt. Dessutom får man upp en sökruta under knappen *Mask*, där man kan söka på ett visst ord, en viss tagg osv. Pilarna bredvid *Mask* gör det dessutom möjligt att klicka sig vidare i sökruvalet. Vi ska tillägga att den siffra som finns mellan pilarna i *Move*-fältet (siffran 400 i figur 2) inte är nuvarande mening utan den senast ändrade meningen.

Fältet *Dependency* innehåller olika alternativ för att bygga träd. Detta fält används kanske främst vid korrigering av trädstrukturer, då man behöver foga samman eller skilja på olika konstituenten, m.m. Slutligen har vi fältet *Edglabel* där man kan godkänna eller byta namn på kantetiketter i trädstrukturen.

### 1.3.3 Annotering steg för steg

Välj den korpus du vill arbeta med i *General*-fältet. Den nya mening du vill arbeta med syns nu i trädfältet, tokeniserad och numrerad på ordnivå. (Om meningen har ordklasstaggats står taggarna under orden. Orden måste taggas innan de parsas.)

Då är vi redo att börja parsa meningen. Genom att högerklicka på musen får man upp parserns första förslag. Om man accepterar förslaget är det bara att gå vidare. Om man anser att förslaget är felaktigt kan man korrigera strukturen med hjälp av olika kommandon. Det finns två sätt att utföra kommandon i *Annotate*. Dels kan man välja kommandon från listan under *Dependency*. Om man t ex vill ta bort en nod markerar man den med musen (den blir vit mot den grå bakgrunden) och väljer alternativet *Ungroup*. Dels kan man använda sig enbart av musen för att utföra alla kommandon. Tabell 1<sup>2</sup> visar hur musknapparna används.

Ibland är parsern inte säker på sitt förslag, utan vill att annoteraren kontrollerar en nod eller etikett. I figur 2 har parsern gett förslag på att *tänker du* är en sats. Detta är korrekt (vi infogar bisatsen senare) och vi bekräftar att *du* är subjekt genom att klicka på *End* i *Edglabel*-fältet. Hade ordet nu inte varit ett subjekt hade vi fått välja en annan etikett i listan för *Edglabel* (om parserna ger flera förslag sorteras dessa i probabilitetsordning). Hade vi helt förkastat förslaget om sats hade vi klickat på *Cancel* i *Edglabel*-fältet.

Om parserna inte ger rätt förslag kan man manuellt välja vad man vill binda ihop. Det spelar här ingen roll i vilken ordning man markerar konstituenterna. Om man därefter högerklickar på musen igen kan man få ett förslag för just de valda konstituenterna av parsern. Om man slutligen vill infoga en konstituent i en annan görs detta genom att markera noderna och välja *Add to*. Här är det viktigt att man först markerar den nod som något ska infogas i och sedan den nod som ska infogas. Om man vill hantera noder utan att använda musen kan man skriva in deras nummer under *Selection*. Noderna skrivs då som en lista separerad av komma eller punkt, intervall markeras med bindestreck.

---

<sup>2</sup>Tabellen är en översättning av tabellen i [Plaehn 1998]

Musknapp	Klick	Villkor	Resultat
Vänster	Dubbel	Pekaren över PoS-tag	Kommando "Tag"
		Pekaren över ordkommentar	Kommando "Comment"
		Pekaren över morfologisk information	Kommando "Morphological information"
		Pekaren över Nodetikett	Kommando "Parentlabel"
		Pekaren över kantetikett	Kommando "Edgelabel"
	Enkel	Håll musknappen nedtryckt	Väljer noder, tidigare valda noder raderas
		Håll musknappen och <Shift> nedtryckt	Utökar valet av noder
Mitt	Enkel	Håll musknappen nedtryckt och flytta pekaren	Den synliga delen av trädfältet förskjuts
		Annars	Kommando "Ungroup"
Höger	Enkel	Pekaren över inte vald nod	Kommando "Add to"
		Alla valda noder är sammanbundna	Kommando "Remove"
		Annars	Kommando "Group"

Tabell 1: Översikt över muskommandon.

## 2 Noderna

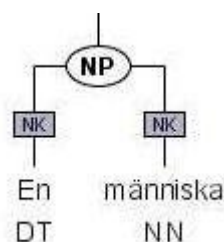
Nedan följer en sammanställning av de vanligaste nodtyperna, hur de används, vilka kantetiketter de kan ha, samt beskrivningen av hur några problemfall bör hanteras. Den som vill ha mer information kan titta i den tyska manualen (ännu inte publicerad), t ex [Brants et al. 2000, Albert et al. 2003].

### 2.1 Nominalfraser

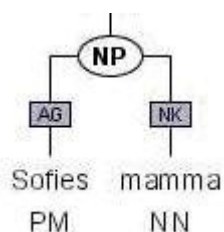
Nod	NP		
Kant	AG	Attributive, Genitive	Genitivattribut
	APP	APPosition	Apposition
	CC	Comparative Complement	Jämförelsekomplement
	CM	CoMparative conjunction	Jämförelsekonjunktion
	MNR	Modifier of Np to the Right	Efterställd NP-modifierare
	MO	MOdifier	Modifierare
	NG	NeGation	Negation
	NK	Noun Kernel	Element i kärn-NP
	OC	Object Clausal	Klausalt objekt
	PAR	PARentthesis	Parentes
	RC	Relative Clause	Relativsats

Noder under NP utelämnas om de enkelt kan rekonstrueras. Om NP:n innehåller ett adjektiv, som alltså utgör en egen AP, markeras denna ej. Om det däremot är en komplex AP eller ADVP så måste denna givetvis markeras. Samma princip ger också att enkla NP:er inte markeras, t ex pronomen, namn eller substantiv som ensamstående fungerar som subjekt eller objekt i en sats.

Ett ensamt substantiv markeras inte som NP, däremot om det har någon form av bestämning, som här obestämd artikel. (NP:n markeras dock inte om den ingår i en PP, se nedan.) Alla delar markeras som kärnelement, NK.

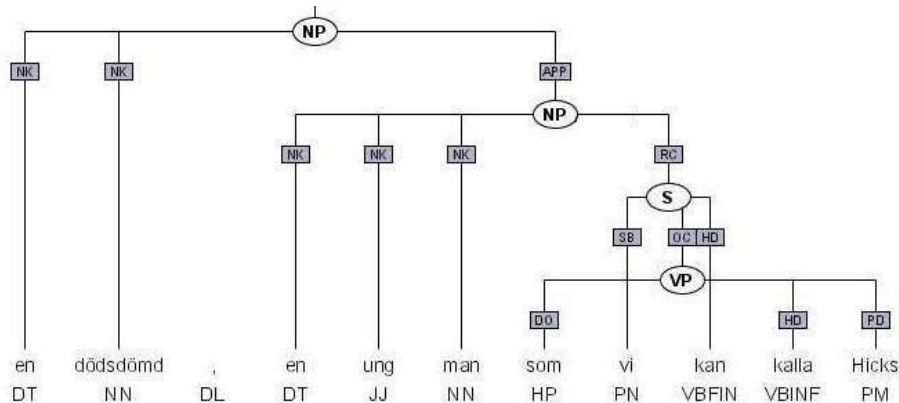


Denna NP innehåller två substantiv. Det ena (huvudet) är NK och det andra ett genitivattribut, AG.

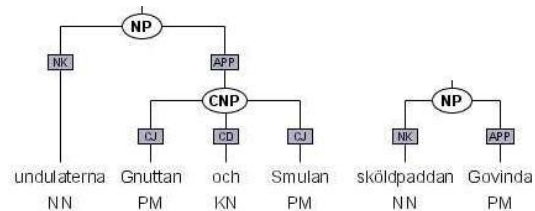




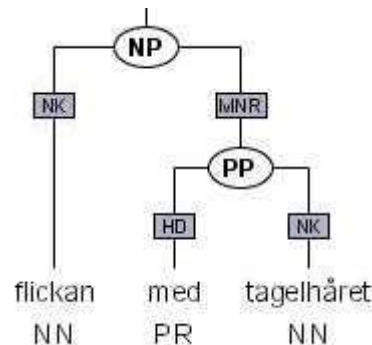
En appositionen (APP) förklarar, definierar eller modifierar det ord/uttryck det står i apposition till. Båda enheterna har samma syntaktiska relation till resten av satsen. Appositionen ska kunna stå istället för det uttryck den definierar, utan att satsen blir ogrammatisk eller får ett annat innehåll.



I dessa två exempel ses namnet (eller namnen i den koordinerade NP:n) som en snävare definition av substantivet. Namnet kan stå i stället för substantivet och markeras därför som APP.

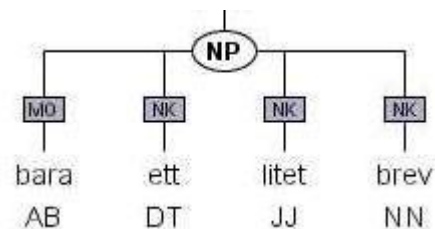


MNR är en efterställd bestämning till en NP, ofta en prepositionsfras. MNR bör skiljas från APP. APP kan stå istället för huvudordet, vilket MNR inte gör.

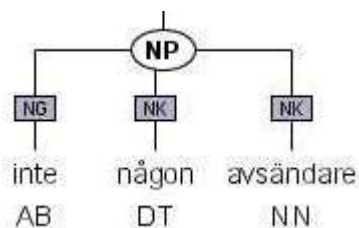


Parentetiska inskott markeras som PAR. De kan vara markerade med parenteser i meningen, "många studenter (*bland dessa befann sig även Axel och Bengt*)", men måste inte vara det, "i dag, *5 år efter den stora hiten*, kommer ingen ihåg honom".

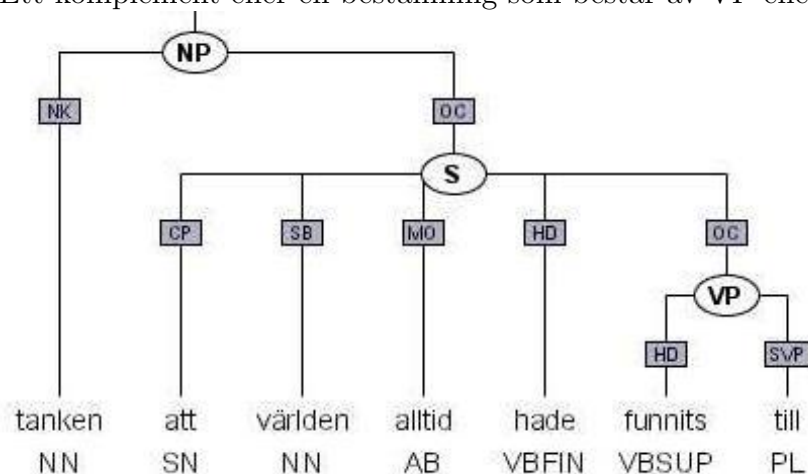
Denna NP innehåller en MO. Adverbi-ala fokuspartiklar (bara, även, åtminstone) annoteras som MO om de semantiskt syftar på NP:n, oavsett placering i NP:n.



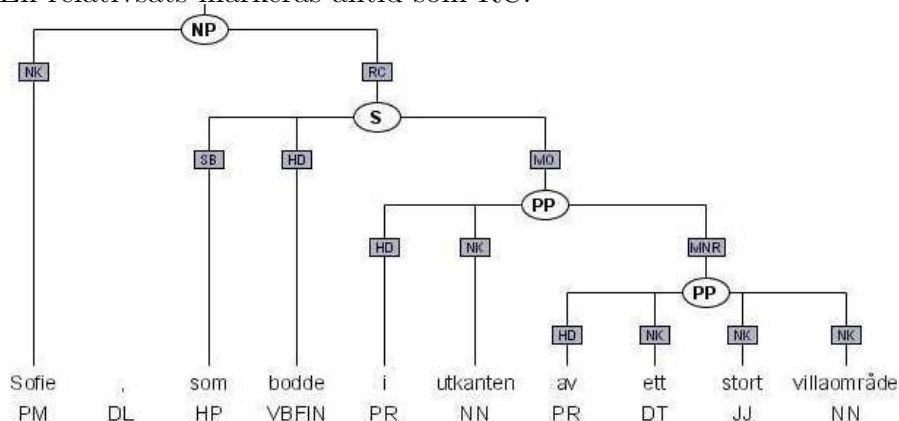
Negationer markeras som NG.



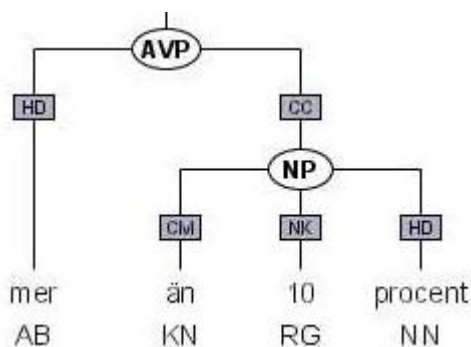
Ett komplement eller en bestämning som består av VP eller S annoteras som OC.



En relativsats markeras alltid som RC.



Ord som “än” och “som” i fraser som “större än” och “lika långt som” markeras som CM i sin NP, PP m.m. Frasen med jämförelsekonjunktionen markeras som jämförelse (CC) i modernoden.

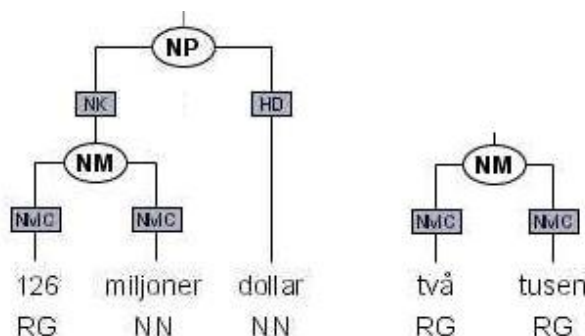


Ordet “som” är problematiskt. I enlighet med SUC-korpusen kan ordet ha tre olika ordklassstagar. Dessa är HP (relativpronomen), KN (konjunktion, särskilt när ordet

fungerar som komparativ konjunktion) samt SN (subjunktion, inleder en underordnad sats, skiljer sig därmed från SUC). I de fall “som” används som komparativ konjunktion är den annoterad som CM i sin NP eller PP. I SUC förekommer dessutom taggen HA (interrogativt/relativt adverb), vilken vi dock hittills inte använt.

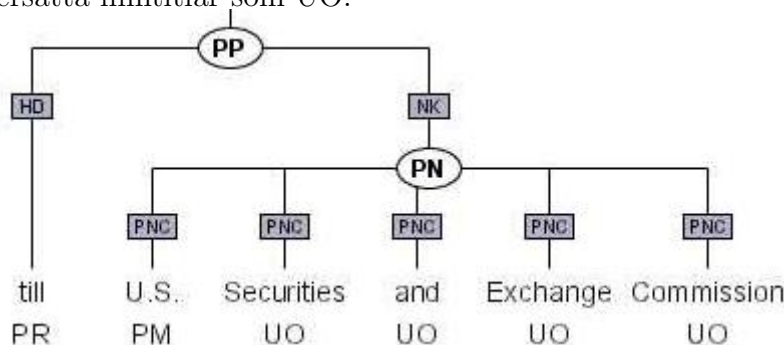
Enkla numeralier har taggen RG och annoteras som NP. Numeralier som “10 000” bör hanteras i förbehandlingen så att de är en enhet, en ordförekomst. Modifierare, som “nästan”, “runt”, annoteras som MO i AP, se kapitel. 2.2.

Komplexa numeralier annoteras med Nodetiketten NM (NuMber), där alla pendants får kantetiketten NMC (NuMber Component).

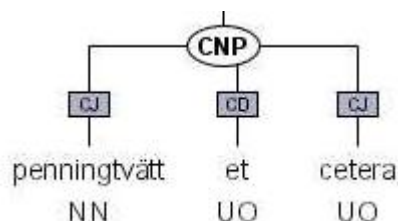


Vid datum och tidsangivelser markeras varje led som NK i NP/PP, “den 12 maj 1975”, “klockan 13”.

Det kan vara svårt att hantera ord och uttryck på främmande språk i en text. Främmande ord ges taggen UO. Hela citat markeras med noden CH (CHunk), där varje pendant får etiketten UC (Unit Component). Observera dock att ord som får anses vara en del av svenska språket inte markeras som UO. Gränsdragningen är här flytande, men vi rekommenderar att allt som kan ses som ett namn (företag, produkter, institutioner m.m.) annoteras som namn, medan sådant som kan/brukar översättas markeras som FM. Det betyder t ex att vi, i motsats till den tyska annoteringsvägledningen, markerar översatta filmtitlar som UO.

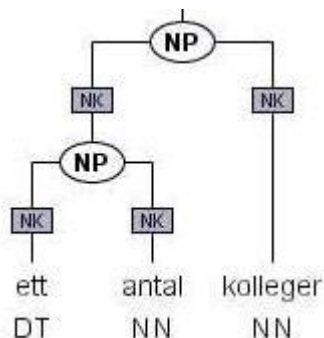


I somliga fall kan uttrycket som helhet anses vara införlivat i svenska språket, men orden taggas ändå som främmande var för sig.

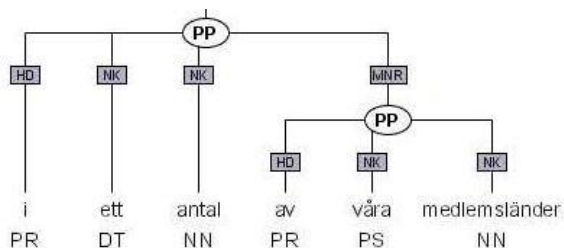


Som sagt markeras bestämmningar till ett ord eller uttryck som pendants till den fras huvudordet ingår i. Det finns dock vissa problem med detta.

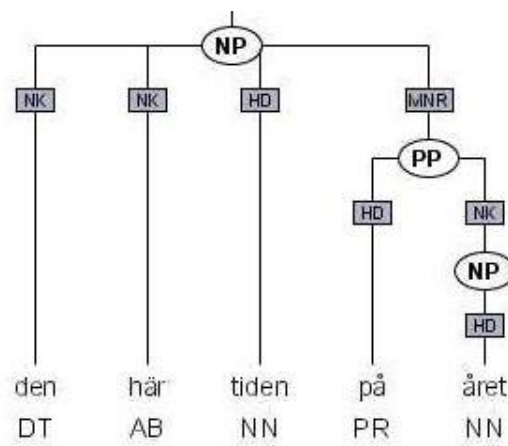
Här markeras bestämningen “ett antal” som pendant till huvudordet “kolleger”.



Eftersom huvudordet “medlemsländer” här föregås av en preposition, kan vi inte markera bestämningen “ett antal” som pendant till huvudordet. Det hamnar istället ovanför sitt huvudord i trädstrukturen.



Ett problem som har lämnats öppet är hur man bör hantera uttryck som “den här”. Vi har låtit orden vara direkta pendants till NP:n.

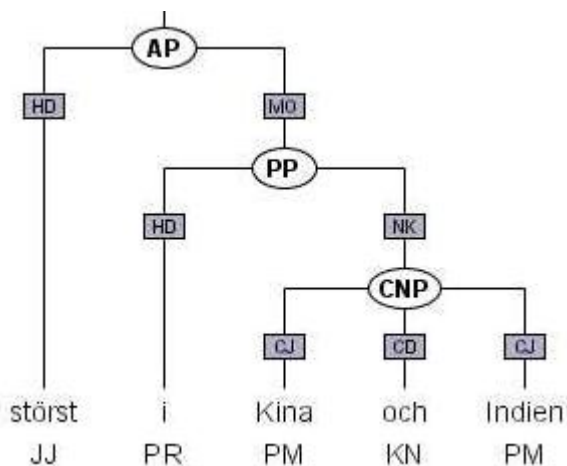


## 2.2 Adjektivfraser

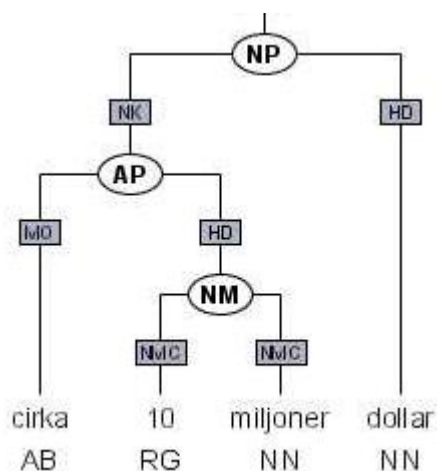
Nod	AP	
Kant	ADC	ADjective Component
	AMS	Measure Argument of Adjective
	CC	Comparative Complement
	CM	CoMparative conjunction
	HD	HeaD
	MO	MOdifier
	NG	NeGation
		Del av komplext adjektiv
		Måttargument till adjektiv
		Jämförelsekomplement
		Jämförelsekonjunktion
		Huvud
		Modifierare
		Negation

I en adjektivfras markeras adjektivet alltid som HD. Två adjektivfraser direkt efter varandra bör undvikas. Om två adjektiv följer på varandra i samma NP annoteras de antingen som en AP, om det ena adjektivet kan ses som modifierare (MO) till det andra (HD), eller som en koordinerad AP, CAP.

Prepositionsfraser och adverb annoteras som MO i AP:n.

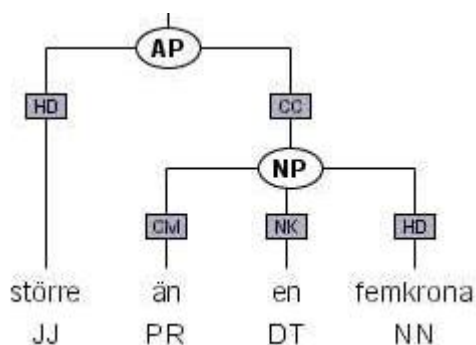


En modifierare (MO) och ett numeriskt uttryck bildar tillsammans en AP, som i sin tur fästs i NP:n.

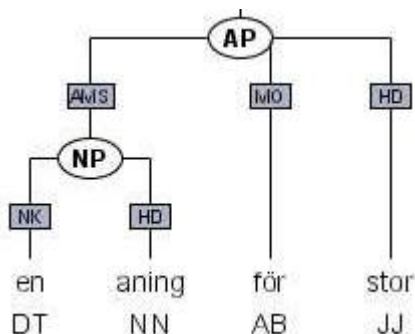


Om ett adjektiv är flerledat (vilket främst förekommer om det är avlett från egenamn) markeras det som MTA (Multi-Token Adjective) i stället för AP. Alla led i frasen markeras då som ADC (ADjective Component). Detta gäller t ex i fall som “den *New Yorkska* pressen”.

Ett adjektiv i komparativ kan bindas ihop med ytterligare ett argument, en fras inledd av “än”. Argumentet får etiketten CC och “än” markeras som CM. Samma sak gäller för adjektivfraser med att argument inlett av “som”.



Etiketten AMS används för uttryck av typen “*två år gammal*”, “*6 procent lägre*” m.m.

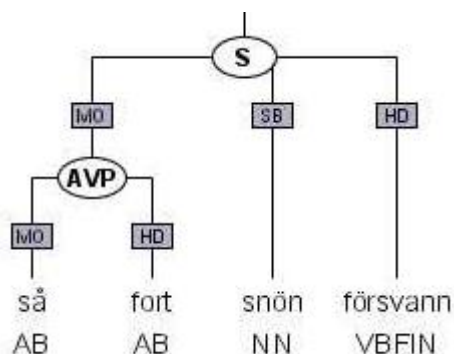


### 2.3 Adverbialfraser

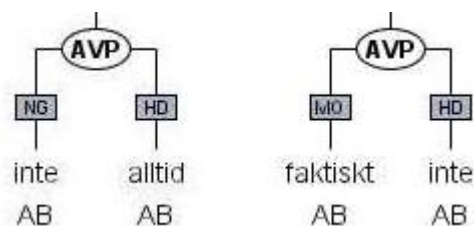
Nod	AVP		
Kant	CC	Comparative Complement	Jämförelsekomplement
	CM	CoMparative conjunction	Jämförelsekonjunktion
	HD	HeaD	Huvud
	MO	MOdifier	Modifierare
	NG	NeGation	Negation

Adverbfraser hanteras på liknande sätt som adjektivfraser. Vi markerar alltid HD i en adverbfras. Som vanligt markeras inte unära adverbfraser.

En AVP är ofta MO i en annan fras. Adverb som modifierar adverb är också MO.



Negationer markeras normalt som NG. En negation kan dock även vara huvud i frasen. Då markeras hela AVP:n som NG i modernoden.



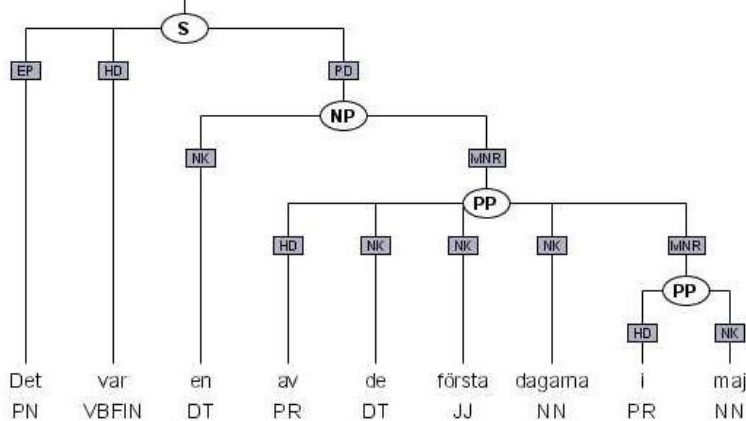
### 2.4 Namn

Namn markeras som PN (Proper Noun) och alla pendants får kantetiketten PNC (Proper Noun Component). PN blir en NK-komponent i NP.

Om ett namn inte har taggats som PM (proper noun), t ex “scoutgruppen *Björnarna*” markeras den alltid som PN, även om det är en unär nod. Observera att detta alltså är ett undantag till regeln om att inte markera unära noder.

## 2.5 Platshållarfraser

I den tyska annoteringsanvisningen finns ett flertal olika typer av platshållarfraser uppräknade. Endast en del av dessa är intressanta för svenska. När “det” kan ersättas med “han” är det ett vanligt pronomen och ska annoteras på samma sätt som NP. Expletivt “det” kan inte ersättas med något annat pronomen, men är obligatoriskt. Typexemplet är “det regnar”. Expletivt “det” markeras som EP i sin fras.

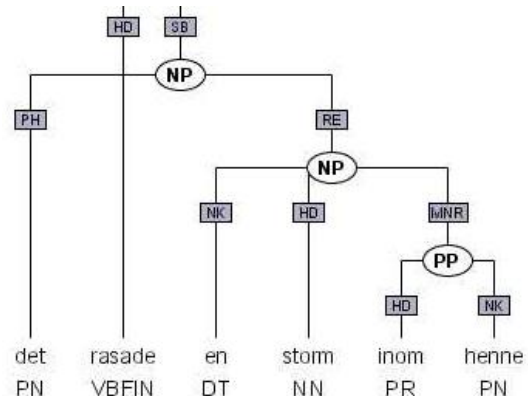


Om “det” inte tillhör någon av de två nämnda kategorierna står det tillsammans med ett subjekt eller objekt. Normalt är detta “det” optionellt.

“I dag låg det bara ett brev i brevlådan.” ⇒ “I dag låg bara ett brev i brevlådan.”

“Det var ganska konstigt att hon fanns till.” ⇒ “Att hon fanns till var ganska konstigt.”

Platshållande “det” markeras som PH (platshållare) och det egentliga argumentet som RE (repeated element).



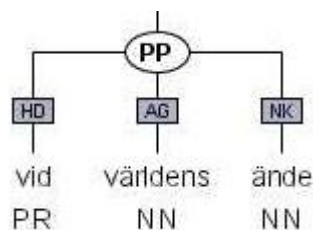
## 2.6 Prepositionsfraser

Nod	PP		
Kant	AG	Attributive, Genitive	Genitivattribut
	APP	APPosition	Apposition
	CC	Comparative Complement	Jämförelsekomplement
	CM	CoMparative conjunction	Jämförelsekonjunktion
	HD	Head	Huvud
	MNR	Modifier of Np to the Right	Efterställd NP-modifierare
	MO	MOdifier	Modifierare
	NG	NeGation	Negation
	NK	Noun Kernel	Element i kärn-NP
	OC	Object Clausal	Klausalt objekt
	PAR	PARentthesis	Parentes
	RC	Relative Clause	Relativsats

Det är viktigt att man skiljer mellan prepositioner och verbpartiklar (se även under verbfraser, s. 17). En generell grundregel är att prepositioner är obetonade, medan partiklar är betonade.

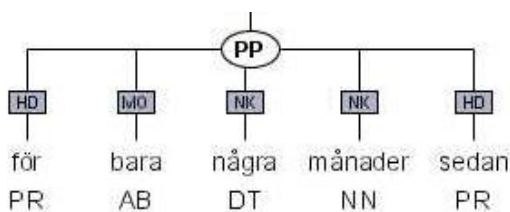
En PP är, precis som NP:n, flat och behandlas även i övrigt precis som en NP. Vi kan se det som att vi flyttar upp nivån för hur allt binds ihop ett steg om vi har en PP jämfört med om vi har en NP. Om allt utom prepositionen ingår i samma NP markeras denna alltså inte inuti PP:n. Prepositionen markeras som HD. (Enligt de tyska anvisningarna markeras den som AC (Adpositional Casemarker), då prepositioner styr kasus i tyskan.) Observera att huvudordet i nominalfrasen inuti PP:n alltså inte markeras som huvud utan som NK.

Denna PP innehåller två substantiv som tillsammans skulle utgöra en NP, men är direkta pendants till PP:n och därför inte markeras.



Tyskan har både pre-, post- och cirkumpositioner. I svenskan är prepositioner vanligast, men det finns även cirkumpositioner.

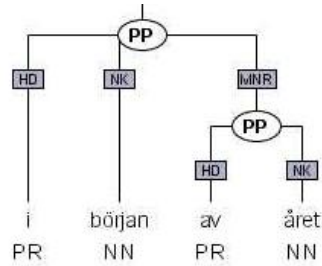
Frasen “för...sedan” är ett exempel på en cirkumposition. Ofta förekommer cirkumpositioner i fasta uttryck, som “i brist på”, “i enlighet med” och “i stället för”.



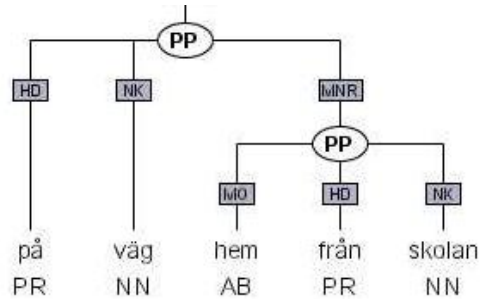
Observera att exempel som “från och med” och “i och med” är en koordinerad prepositionsfras, CAC (se kapitel 2.8), som i sin tur fungerar som preposition i prepositionsfrasen, PP.



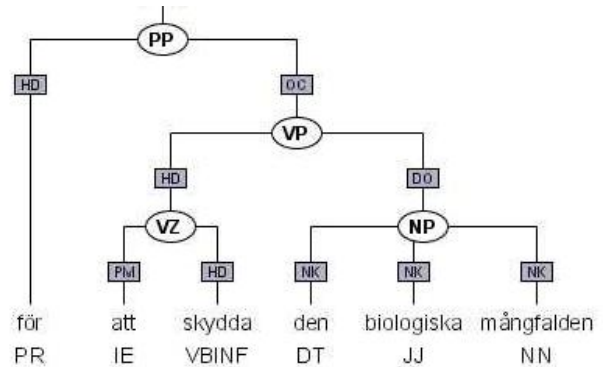
I en PP är en MNR ofta en annan PP.



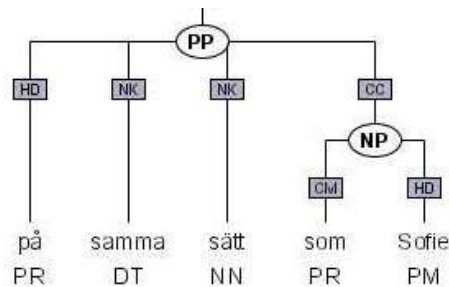
Ibland kan det vara svårt att avgöra vilken nod ett ord ska tillhöra. Här skulle man kunna argumentera för att placera ordet "hem" både i den övre och den undre PP:n. Det viktiga är att man försöker vara konsekvent. Är man riktigt osäker är det alltid bättre att binda en pendant högt än lågt.



På svenska kan rektionen i en PP utgöras av en sats eller verbfras, vilken markeras som OC.



Jämförelsekomplement markeras som CC, här ett exempel på en NP som är CC i en PP.

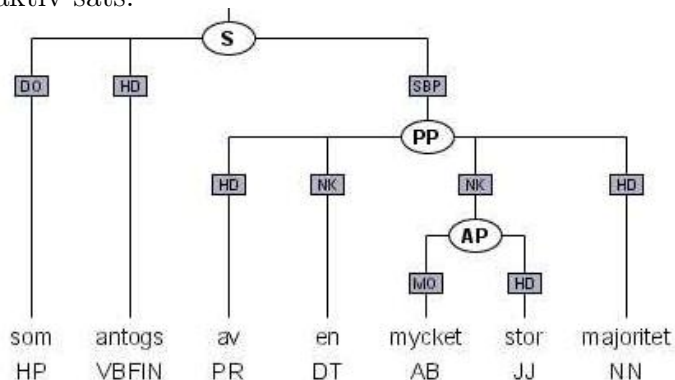


## 2.7 Satser och verbfraser

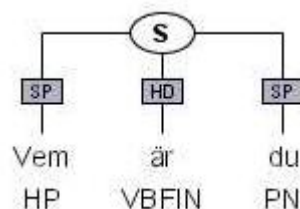
Nod	S, VP		
Kant	CP	ComPlementizer	Subjunktion
	DM	Discourse Marker	Diskursmarkör
	HD	HeaD	Huvud
	JU	JUjctor	Konjunktion
	MO	MOdifier	Modifierare
	NG	NeGation	Negation
	OC	Object Clausal	Klausalt objekt
	DO	Direct Object	Direkt objekt
	IO	Indirect Object	Indirekt objekt
	OP	Object Prepositional	Prepositionsobjekt
	PD	PreDicative	Predikativ
	SB	SuBject	Subjekt
	SBP	SuBject Passivised	Passivt subjekt
	SP	Subject or Predicative	Subjekt eller predikativ
	SVP	Special Verb Particle	Verbpartikel
	VO	VOcative	Vokativ

Varje verb blir HD i sin egen fras. S och VP har liknande användning och kan i stort innehålla samma typer av kanter och dotternoder. Skillnaden mellan dessa två är att S används för finita satser, medan VP används för infinita satser. Tack vare den flata strukturen markeras alltså inte verbet med rektion som VP i en finit sats.

Subjektet är alltid dependent till det finita verbet och binds alltså till S-noden med kantetiketten SB. SBP används i passivkonstruktioner, för det som hade varit subjekt i en aktiv sats.



SP används när det är svårt att avgöra vilket som är subjekt och vilket som är predikativ i en sats.

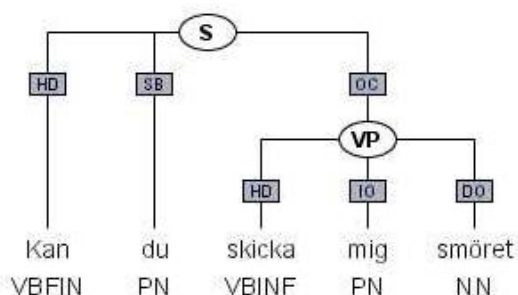


Som predikativ markeras endast NP- och AP-fraser (med eller utan nod) vid verben

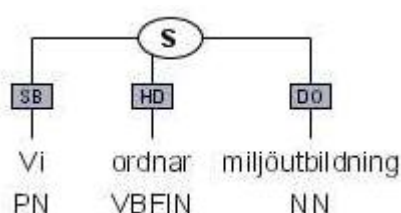
vara, bli, heta, kallas. Kopulan annoteras som HD i sin S/VP.

Det finns flera olika typer av objekt. I de tyska anvisningarna förekommer ackusativ-, dativ- och genitivobjekt. I svenskan talar vi istället om indirekt och direkt objekt.

Indirekta objekt förekommer endast vid bitransitiva verb och betecknar det objekt som står närmast verbet eller styrs av prepositionen.

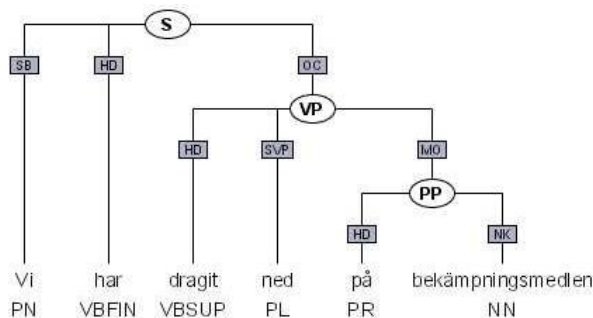


Direkt objekt betecknar det återstående objektet vid bitransitiva verb samt objektet vid monotransitiva verb.

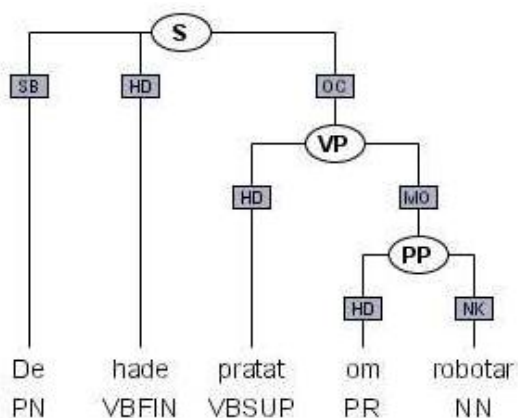


Verbpartikeln taggas som PL och får kantetiketten SVP. (SVP står enligt den tyska annoteringsanvisningen för Separable Verb Prefix. För svenska har vi döpt om kanten till Special Verb Particle.) Det kan vara svårt att skilja mellan en verbpartikel och en preposition. Generellt sett gäller att verbpartikeln står i nära relation till verbet och är betonad, i motsats till den obetonade prepositionen. I motsats till adverb, som följer på objektet, föregår partikeln ett eventuellt objekt. (Se även exempel under prepositionsfraser, s. 14.)

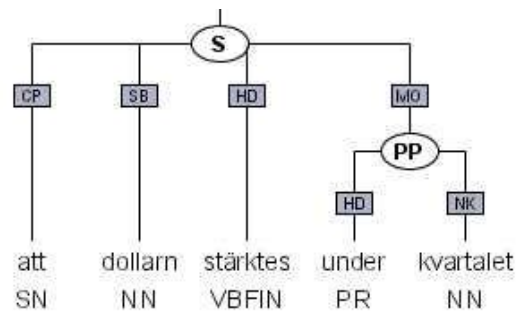
Det är inte ovanligt att en verbpartikel föregår en prepositionsfras.



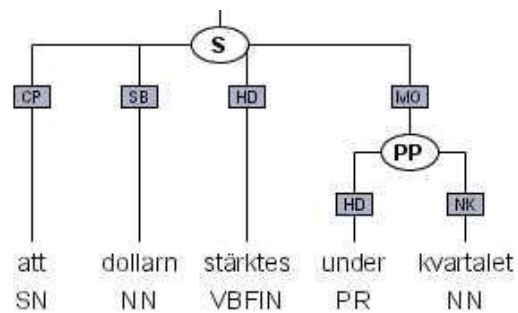
Satskomplement markeras som OC.



Alla subjunktioner (SN) markeras som CP...



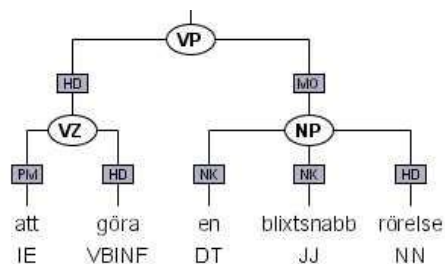
... medan alla konjunktioner (KN) markeras som JU.



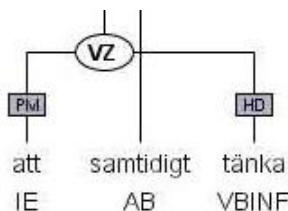
Svarsord och liknande markeras som DM, medan tilltal markeras som VO.

Ett verb i infinitiv tillsammans med sitt infinitivmärke markeras med nodetiketten VZ. I VZ ingår endast verb och infinitivmärke, alla verbkomplement och verbadjunkter placeras i VP-noden. På tyska är infinitivmärket ("zu", därav nodens namn) starkare knutet till sitt verb än på svenska, eftersom ingenting kan placeras mellan dem. Vi valde att använda denna nodetikett även för svenska, även om det kan diskuteras hur relevant detta egentligen är.

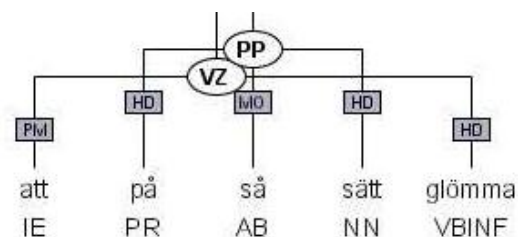
En VZ med infinitiv och infinitivmärke.



I VZ kan ord placeras mellan infinitivmärket och verbet, vilket leder till att vi måste använda korsande kanter. Adverbet blir MO i VP:n.



Här har en hel PP skjutits in mellan infinitivmärket och verbet. Även denna blir MO i VP:n.



Ibland förekommer satser utan verb. Detta kan bero på att satsen är ofullständig (t ex rubriker) eller att hjälp verbet “ha” har utelämnats. Om satsen är ofullständig bör den annoteras som om den hade varit fullständig. Om det är svårt att rekonstruera den fullständiga meningen är det ofta bättre att lämna delarna som fraser eller fragment utan att binda ihop dem.

I svenskan kan hjälp verbet “ha” utelämnas om det är ett infinit led i en verbkedja, om det ingår i en huvudsats som innehåller ordet “kanske”, samt om det ingår i en bisats där verbfrasens tempus är perfekt eller pluskvamperfekt. I somliga fall får vi alltså ett S utan huvud. I övriga fall måste man bestämma om man vill markera en tom VP för hjälp verbet eller om man ignorerar att ett led saknas och inte markerar det alls.

“Allt måste (ha) haft en begynnelse.”

“Hon kanske (hade) kunnat välja sina vänner.”

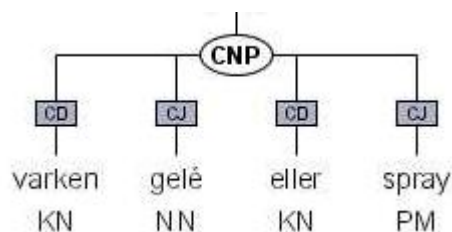
“Jag betvivlar inte att kammaren (har) haft möjlighet att granska varje ändringsförslag.”

“hönsen, som under kriget (hade) fått gå omkring i trädgården”

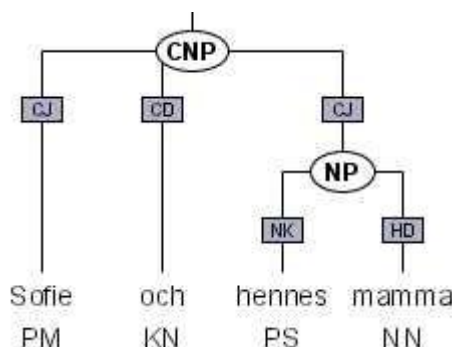
## 2.8 Koordinerade fraser

Koordinerade fraser består av en eller fler konjunkter (CJ), som kan bindas samman av koordinerande konjunktioner (CD). Den koordinerade frasen har samma namn (med framförstående C) som de delar som ingår i den. Två NP som sammanbinds av en konjunktion bildar alltså en CNP. Vidare förekommer CAP, CAVP, CPP, CS, CVP och CVZ. Dessutom finns några typer av koordinerade fraser, vars delar i sig inte räknas som egna fraser, utan normalt markeras med kantetiketter. Dessa är CAC (coordinated adposition) och CCP (coordinated complementizer).

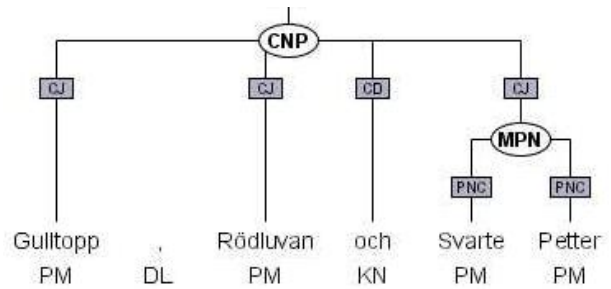
En CNP bestående av två substantiv (som, då de är enkla noder, inte får någon egen NP-nod) samt två konjunktioner.



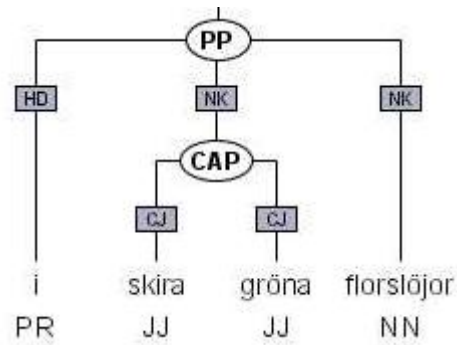
En CNP, där ena konjunkten är komplex NP och den andra är ett ensamt substantiv.



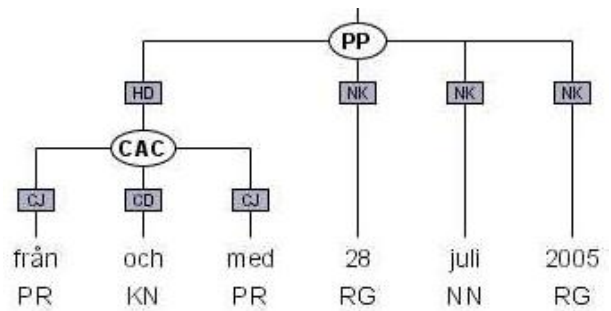
Konjunkterna kan separeras med konjunktioner eller skiljetecken.



Det förekommer att en koordinerad fras saknar sammanbindande konjunktion eller skiljetecken. Inget av dessa två adjektiv är en bestämning till det andra, varför det markeras som CAP.



En CAC är koordinerade prepositioner, och noden markeras som HD i PP:n.



## Referenser

- [Albert et al. 2003] Stefanie Albert, Jan Anderssen, Regine Bader, Stephanie Becker, Tobias Bracht, Sabine Brants, Thorsten Brants, Vera Demberg, Stefanie Dipper, Peter Eisenberg, Silvia Hansen, Hagen Hirschmann, Juliane Janitzek, Carolin Kirstein, Robert Langner, Lukas Michelbacher, Oliver Plaehn, Cordula Preis, Marcus Pussel, Marco Rower, Bettina Schrader, Anne Schwartz, George Smith, and Hans Uszkoreit. 2003. TIGER Annotationsschema. July.
- [Brants et al. 2000] Thorsten Brants, Stefanie Dipper, Peter Eisenberg, Sabine Kramp, Cordula Preis, Marcus Pussel, Anne Schwartz, George Smith, and Hans Uszkoreit. 2000. TIGER Annotationsschema. May.
- [Plaehn 1998] Oliver Plaehn, 1998. *ANNOTATE - Bedienungsanleitung*.
- [Samuelsson and Volk 2004] Yvonne Samuelsson and Martin Volk. 2004. Automatic node insertion for treebank deepening. In *Proceedings of the 3rd Workshop on Treebanks and Linguistic Theories (TLT2004)*, Tübingen, Germany, December.
- [Samuelsson 2004] Yvonne Samuelsson. 2004. Parallel Phrases - Going Automatic. Further experiments towards a German-Swedish parallel treebank. Master's thesis, Stockholm University.
- [Volk and Samuelsson 2004] Martin Volk and Yvonne Samuelsson. 2004. Bootstrapping parallel treebanks. In *Proceedings of the 5th International Workshop on linguistically Interpreted Corpora (COLING2004)*, pages 63–69, Geneva, Switzerland, August.

## A SUC ordklasstaggar

Observera att denna lista skiljer sig något från den “officiella” listan över SUC-taggar. Vi använder bara de rena ordklasstaggar, inte de morfologiska taggarna. Taggar markerade med \* har lagts till.

SUC-tagg	Beskrivning
AB	Adverb
DL	Delimiter
DT	Determiner
HA	Interrog./rel. adverb, wh-adverb
HD	Interrog./rel. determiner, wh-determiner
HP	Interrog./rel. pronoun, wh-pronoun
HS	Interrog./rel. possessive, wh-possessive
IE	Infinitive marker
IN	Interjection
JJ	Adjective
KN	Conjunction
NN	Noun
PC	Participle
PL	Verb particle
PM	Proper noun
PN	Pronoun
PR	Preposition
PS	Possessive
RG	Cardinal number
RO	Ordinal number
SN	Subjunction
UO	Foreign word
VBFIN *	Verb, finite
VBINF *	Verb, infinitive
VBIMP *	Verb, imperative
VBSUP *	Verb, past participle