

UPPSALA UNIVERSITY,
UPPSALA, SWEDEN
COMPUTER SCIENCES DEPARTMENT
JANUARY, 1971

Enkel PCF för svenska

av

Erik Sandewall

Sammanfattning: I tidigare rapporter har vi föreslagit en metod för användning av predikatkalkyl som djupstruktur, varvid djupstrukturen har formen av ett träd på vanligt sätt, men vissa begreppsmässiga primitiver (prepositioner, artiklar, o.dyl.) ligger i trädets förgreningpunkter snarare än i slutnoderna. Ur predikatkalkylens synpunkt är dessa primitiver vanligen funktioner. - I denna rapport ställer vi upp några primitiver som behövs för representation av enkla svenska meningar, och gör testöversättningar. Rapporten är mycket preliminär.

Detta arbete har bedrivits med stöd av Naturvetenskapliga forskningsrådet (anslag Dnr 2654-3)

1. Grundprinciper

I denna rapport följer vi de förutsättningar som gjorts i de tidigare rapporterna "Representing natural-language information in predicate calculus", "Formal methods in the design of question-answering systems", och "Towards a formal theory of discourse". Förutsättningarna kan sammanfattas på följande sätt:

1. Vid analys av en mening är uppgiften att relatera meningen till en given databas, dvs. en samling av tidigare information.
2. Analysen bör gå i följande steg:
 - a. In-översättning. Denna överför den givna meningen till ett formeluttryck eller (ekvivalent) ett träd.
 - b. Assimilering. Härvid ersätts uttryck som inleds med t.ex. bestämd artikel med de noder i databasen som de refererar till.
 - c. Beroende på meningens typ, utföres något av följande:
 - c1. För påståenden: acceptans. Härvid undersökes om meningen kan accepteras i databasen, och i såfall införes den.
 - c2. För frågor: frågebesvarande. Härvid bestäms svaret på frågan ur databasen.
 Ytterligare alternativ (c3, c4,...) är möjliga.

Med djupstruktur avser vi resultatet av inöversättningen, dvs. steg (a). Denna rapport skall behandla dels innehållet i djupstrukturen, dels inöversättningen som tar en given mening till djupstrukturen. Inöversättningen kommer att diskuteras genom en case-study, nämligen av en barnbok ("När Babar var liten").

Inöversättningen förutsättes mycket schematiskt gå i följande steg:

- aa. Morfologisk och lexikalisk analys. Härvid avskiljes ändelser, och

märkes orden med ordklass. Detta kräver normalt tillgång till lexikon. Om lexikonet är ofullständigt eller en flertydighet föreligger ("anden", "leder") eller talaren använder ett ord som han har hittat på själv, gör systemet en gissning, men sparar möjligheten av senare bakspårning.

- ab. Strängtransformationer. Dessa är transformationer på den givna meningen, betraktad som en sträng, och tjänar till att göra om ordföljden m.m.
- ac. Infix-prefix-överföring. Efter strängsträngtransformationerna betraktas meningen som ett uttryck uppbyggt med binära infix-operatorer, samt med monära prefix och postfix. Nästa steg är nu att överföra allt till fullständigt parentetiserad prefix, vilket är ekvivalent med att göra om strängen till ett träd.
- ad. Strukturtransformationer. Dessa är transformationer på det erhållna formeluttrycket eller trädet, vilka skall överföra uttrycket till assimilerbar och accepterbar form. Här ingår bl.a. tillägg av det konnektiv varmed en påståendesats förknippas med tidigare satser ("sedan", "och", el.dyl.)

De fyra delstegen (aa) - (ad) är lämpliga som arbetshypotes i denna rapport, men blir troligen opraktiska vid programmering. Om man t.ex. vill använda Wood's parser föredrar man säkert att göra allt på annat sätt. De skall alltså uppfattas som ett pedagogiskt trick.

2. Begreppsmässiga primitiver

I denna sektion införes de operatorer som skall stå i förgreningspunkterna av djupstrukturträdet, eller (ekvivalent) de funktioner m.m. som behövs i predikatkalkylformuleringen. Vi skall illustrera operatorerna med exempel skrivna som formler i infixform, dvs. den form meningen har efter strängtransformationerna men före infix-prefix-överföringen.

Först vissa notationskonventioner:

Olika sorters operatorer. Vi skriver t.ex.

peter	konstant (objekt)
far	konstant (egenskap)
far TIL lars	"TIL" är infix
En pojke	"En" är prefix
pojke -n	"-n" är postfix
Mer- stor -ÄN- lars	"Mer- ... -ÄN-" är bifix

Relationer m.m. Relationer ur predikatkalkylens synpunkt skrives kursiverade. Exempel:

peter *ÄR* far TIL lars

Med dessa konventioner kan vi nu bäst ange de olika primitiverna i tabellform. I givna exempel anges infixformen inom citationstecken och naturligt-språk-formen inom parentes. Parenteser anges i infixformen i den mån de inte är triviala. - Operatorerna indelas i kategorier efter sitt användningssätt.

Översikt över primitiverna

<u>Kategori</u>	<u>Primitiv</u>	<u>Kommentar</u>
Satsbildare	ÄR	motsv. WERE resp. IZ (vi kör inte med IS, utan sätter <i>Satsbild</i> före hela satsen)
	<i>Santatt</i>	motsv. Istrue
	OBJ	binder ihop verb och akkusativobjekt
	TIH	binder ihop resultatet med dativobjekt
	SÅ	binder ihop verb el dyl med satsadverbial: i satsen "Han cyklar ofta" anses schematiskt konnektivet SÅ stå före "ofta".
Artiklar (Elimineras vid assimiler- ing)	-n, Den (x)	idealiserad bestämd artikel
	En, Ett	idealiserad obestämd artikel (existenskvant.)
	Varje, -or, -ar	allkvantifierare
	-orna, -arna, -erna	idealiserad bestämd artikel som hänvisar till tidigare konglomerat, bildat med ST (jfr nedan)
Bildare av nominal- uttryck	OCK	"grå OCK elefant"
	BL	"liten BL elefant" (<u>b</u> land)
	SS	"dålig SS lärare" (<u>så</u> som)
	BOS	ersätter vilken som helst av BL, OCK, SS. Anses stå underförstådd mellan två adjektiv och substantiv
	'S	genitiv (peter 'S far)
	TIL	som 'S men motsatt ordning på argumenten (far TIL peter)
	MED	"hus MED tak"
	ST	bildar konglomerat: "sjutton ST bil", "många ST bil"
	Andra- -ÄN-	"Andra- elefant -ÄN- babar" (avbildar egenskap gånger objekt på egenskap)

(x) När flera primitiver upprepas på samma rad uppfattas de som synonyma

Pronomen han, honom
(idealiserade) hon, henne
(elimineras den
vid assimiler- det
ing) de, dem

Refererar till tidigare nämnt objekt med rätt genus etc. Reglerna för i vilken ordning man skall pröva referens är trassliga och får byggas ut allteftersom, men här är några huvuddrag:

- (1) objekt i första ledet av ett OCK-uttryck eller ELLER-uttryck där vi nu analyserar det andra ledet: "Babar och hans mamma"
- (2) subjektet i föregående sats
- (3) akkusativobjekt i föregående sats etc. Vi har inte referens till subjektet i egen sats ("Peter slår honom"), men vi kan ha till subjektet i tidigare sats inom samma mening ("Peter är stor men han kan inte gå").

sig, sigsjälv

Refererar i en sats till dess subjekt

Hans
Hennes
Dess
Deras

"Hans far" uppfattas som

"Den (han2 'S far)", där "han2" är ett pronomen liksom "han", utom att det om det förekommer senare i satsen än akkusativobjektet pekar tillbaka till detta med prioritet 1 1/2: "Peter ger Lars hans pennkniv".

Sin, Sitt, Sina

"Sin far" uppfattas som "Den (sig 'S far)"

Andra

assimileras till ett Andra- -ÄN- uttryck (jfr ovan)

Prepositioner I

anger position

TILL

anger riktning och uppnående (gå till stan)

MEDAN

tid (inklusion i tidsintervall)

Anm. Så småningom kommer ett stort antal prepositioner att behövas.

Verb-
modifierare

-pres

modifierar verb till presens

-skall

"lärs ÄR äta -skall" (Lars skall äta)

ÅKK

"sitta ÅKK läsa"

Omformare

-t

adjektiv till adverb

Att

verbegenskap (verb med ev bestämmingar) till abstrakt objekt: (försöka (att) läsa)

Allm

egenskap (vanl nominalegenskap) till abstrakt objekt (berätta sagor, se syner)

3. Några vanliga strängtransformationer

I nästa avsnitt kommer vi att ge exempel på meningar i vanlig svenska, omskrivna på vad som väsentligen är infixform (dvs formen efter morfologisk analys och strängtransformationer, men före infix-prefix-överföring). Detta är den lämpligaste presentationsformen, eftersom uttrycken blir betydligt mindre läsbara när de kommer på prefixform, och eftersom överföringen dit ändå är trivial. Det är istället i den morfologiska analysen och strängtransformationerna som de väsentliga svårigheterna ligger.

Emellertid visar det sig att vissa saker i prefixformen gör även denna något svårläst, och det är därför lämpligt att dra fram presentationsstadiet ytterligare något. Det vi presenterar i nästa sektion är därför noga räknat det som blir infixform sedan ytterligare några få, här specificerade strängtransformationer genomförts. Det är speciellt verbens böjning som härigenom blir mer läsbar.

Den första av dessa "resttransformationer" är den som flyttar om presensändelse och dylikt. Vi använder beteckningen -R för en presensändelse, och ger här exempel på några meningars väg. För varje mening anges först naturligt språk, därefter nästan-infix (dvs notationen i nästa sektion), och slutligen infix (dvs efter alla transformationer).

- (a) Peter sitter
- (b) peter sitta -R
- (c) peter ÄR sitta -pres

- (a) Peter ser Lars
- (b) peter se -R lars
- (c) peter ÄR se -pres OBJ lars

- (a) Peter sitter i stolen
- (b) peter sitta -R I stol -en
- (c) peter ÄR sitta -pres I stol -en

Överföringen från (b) till (c) är alltså rätt enkel men inte helt trivial. Ändå besvärligare blir det i exempel som

- (a) Peter ser sällan sin giraff
- (b) peter se -R SÅ sällan Sin giraff
- (c) peter ÄR se -pres SÅ sällan OBJ Sin giraff

Den andra "resttransformationen" tar hand om adverbial i början av meningen. Transformationen beskrivs bäst med exempel:

- (a) När Anders sitter i stolen, går åsken
- (b) NÄR anders sitta -R I stol -en, åska -n gå -R
- (b') åska -n gå -R NÄR (anders sitta -R I stol -en)
- (c) åska -n ÄR gå -pres NÄR (anders ÄR sitta -pres I stol -en)

Transformationerna säger alltså att en prepositionsfras som står först skall flyttas till sist i meningen.

En tredje transformation tar hand om fraser av typen "taket på huset". I infixnotationen skall en sådan fras uttryckas som

Det (tak PÅ hus -et)

eller ekvivalent

(tak PÅ hus -et) -et

Ingendera är särskilt läsbart. Istället skriver vi i nästan-infixnotationen

tak -et PÅ hus -et

och förutsätter att detta transformeras till endera av föregående. En sådan transformation skall göras när bestämd artikel (-en, -et) står direkt före en (idealiserad) "nominalkonjunktion", men inte när det står före en preposition eller en "satskonjunktion". Uppenbarligen aktualiserar detta vissa flertydighetsproblem för inöversättaren (ss. i "Han såg taket på huset" resp "Han satte hatten på huvudet"), men detta får lösas t.ex. genom att acceptance ser felet och kräver en bakspårning. Det är f.ö. mycket möjligt att ovanstående klaras bättre med en strukturtransformation.

Observera att den här införda nästan-infix-notationen är en konstruerad hjälnotation för att göra exemplen i nästa avsnitt mer läsbara. Det är inte självklart att en inöversättare ska gå via nästan-infix-notationen, och även om det gör det, måste svårare meningar tas om hand av andra transformationer innan man kommer fram dit.

4. Demonstrationsöversättning av "När Babar var liten" till PCF i nästan-infix-notation

I följande meningar har vi slarvat med prepositionsvalet. Vi anger (c)-versionen (ren infixform) då den behövs. I nästa sektion ges trädrepresentationen. Denna sektion och sektion 5 behöver därför ev. läsas parallellt ett tag.

1. (a) Djupt inne i den stora skogen bor en liten elefant.
 (b) I djup -et AV Den (stor BOS skog),
 En liten BOS elefant bo -R
 (c) En liten BOS elefant ÄR bo -pres I
 Det (djup AV Den stor BOS skog)
 2. (a) Han heter Babar
 (b) han heta -R babar
 (c) han ÄR heta -pres OBJ babar
 3. (a) Hans mamma tycker mycket om sin lilla elefant
 (b) Hans mamma tycka-om -R SÅ mycket
 Sin (liten BOS elefant)
- Anm. "tycka om" uppfattas alltså som ett verb ihop.
4. (a) När han skall sova, vaggar hon honom till sömns med snabeln
 (b) NÄR han ÄR sova -skall, hon vagga -R honom TILL sömns
 MED snabel -n
 5. (a) När han skall äta, berättar hon sagor för honom.
 (b) NÄR han ÄR äta -skall, hon berätta -R Allm saga
 FÖR honom
 6. (a) Om morgnarna badar hon honom och borstar honom på ryggen.

Anm. Det abstrakta substantivet "morgon" (liksom "kväll", etc.) är rätt trassligt. Det skulle vara enklare om man hade ett verb, "att morgna", som användes i analogi med "att regna". Ovanstående mening

skulle då uttryckas som "När det morgnar, ...". Vi följer den varianten - det väsentliga just nu är att få en notation som på något sätt återuttrycker naturligt-språk-informationen. En formulering som noggrannare följer "På morgnarna..." är naturligtvis möjlig men kräver att man grejar med en särskild sort för tidssegment, osv. Därför:

- (b) NÄR morgna -R, hon (bada -R honom OCH borsta -R honom
På rygg -en)

7. (a) Han lär sig fort att gå, och snart kan han springa efter fjärilarna.

- (b) han lära-sig -R SÅ fort Att gå
OCH han kunna -R SÅ snart Att (springa EFTER
Allm fjärl)

Anm. Det kan diskuteras om "snart" ska uppfattas som "snar -t"

8. (a) Babar växer, och snart är han ute och leker med de andra elefanterna.

- (b) babar växa -R OCH han ÄR (ute ÅKK leka -pres) SÅ
snart MED (Andra elefant) -erna

Anm. "vara ute och leka" uppfattas som en speciell svensk konstruktion, analog med t.ex. "sitta och läsa", och beskrivs med ett par av egenskaper, hopkopplade med en variant av och som vi stavar åkk. Det förutsätts att strängtransformationsdelen av inöversättningen skall vara så flexibel att man lätt kan lägga till kod för att hantera sådana konstruktionsklasser, när man kommer på dem.

9. (a) Till frukost får han bananer och apelsiner

- (b) TILL frukost, han få -R Allm banan OCH Allm apelsin

10. (a) Sedan gräver han i sanden med ett vackert snäckskal

- (b) han gräva -R SÅ sedan I sand -en MED
Ett (vacker BOS snäckskal)

11. (a) Ibland går elefanterna till dammen och badar och duschar

- (b) elefant -erna gå -R SÅ ibland TILL damm -en FÖRATT
(bada OCH duscha)

12. (a) Babar tycker mycket om att leka kurragömma
 (b) babar tycka-om -R SÅ mycket
 Att (leka OBJ kurragömma)

13. (a) Elefanterna har det härligt i den stora skogen
 (b) elefant -erna ha-det -R SÅ härlig -t
 I Den (stor BOS skog)

Anm. Att "ha det" (härligt, roligt, tråkigt,...) uppfattas alltså här som ett intransitivt verb.

14. (a) En dag är Babar och hans mamma ute och går i skogen.
 (b) MEDAN En dag, babar OCH Hans mamma ÄR (ute ÅKK
 gå) -pres I skog -en

15. (a) Babar får rida på sin mammas rygg
 (b) babar få -R Att (rida PÅ (Sin mamma) 'S rygg)

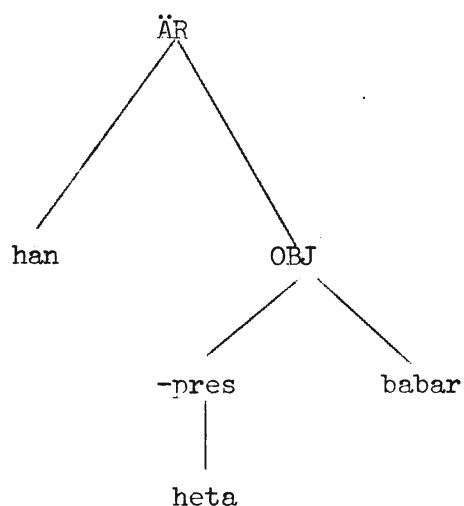
5. Trädrepresentation

Metodiken för att omvandla en (c)-formel i sektion 3 eller 4 till ett träd (P-marker) är enkel:

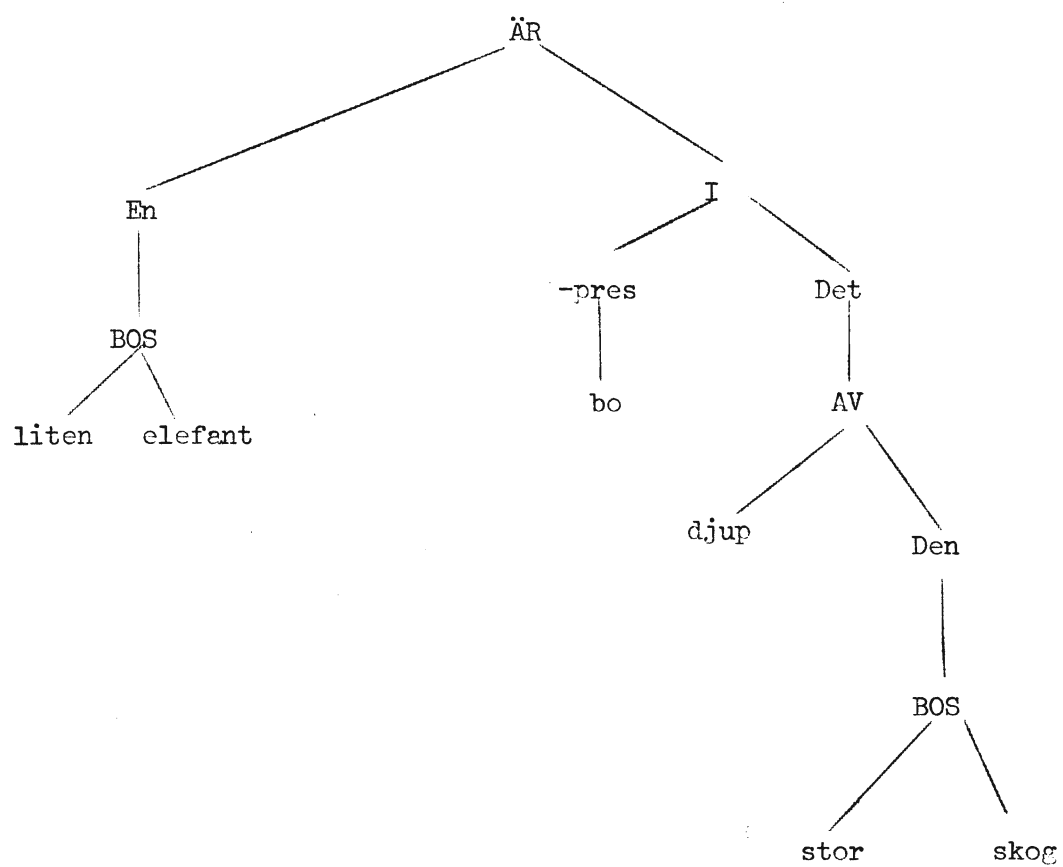
1. Kolla operatorernas aritet (antal argument) och om det är prefix eller postfix
2. Parentetisera korrekt. Detta görs med hjälp av ariteten och med hjälp av förståelsen av vad meningen betyder.
3. Rita upp trädet enligt parentetiseringen med funktionerna i förgreningspunkterna.

Vi skall ge ett par exempel. Numrering som i sektion 4.

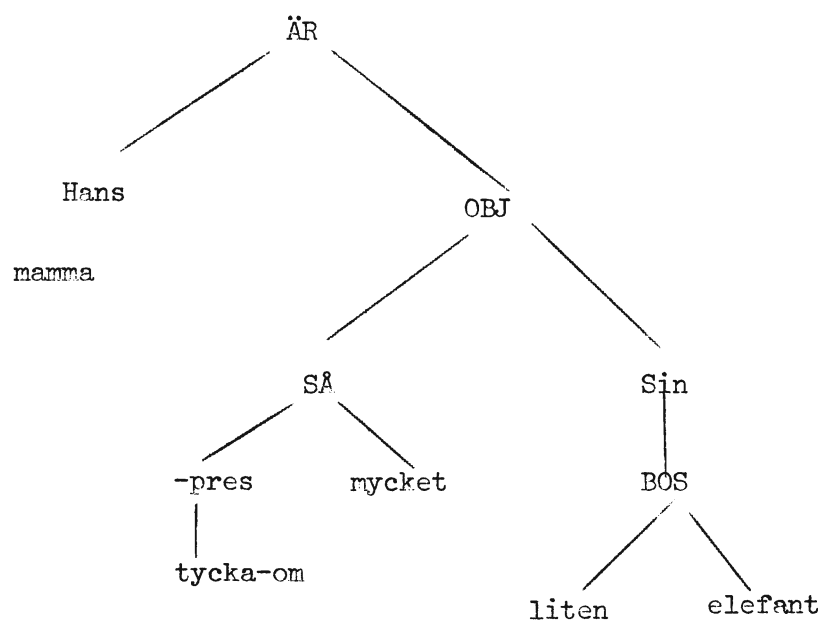
2. han ÄR heta -pres OBJ babar



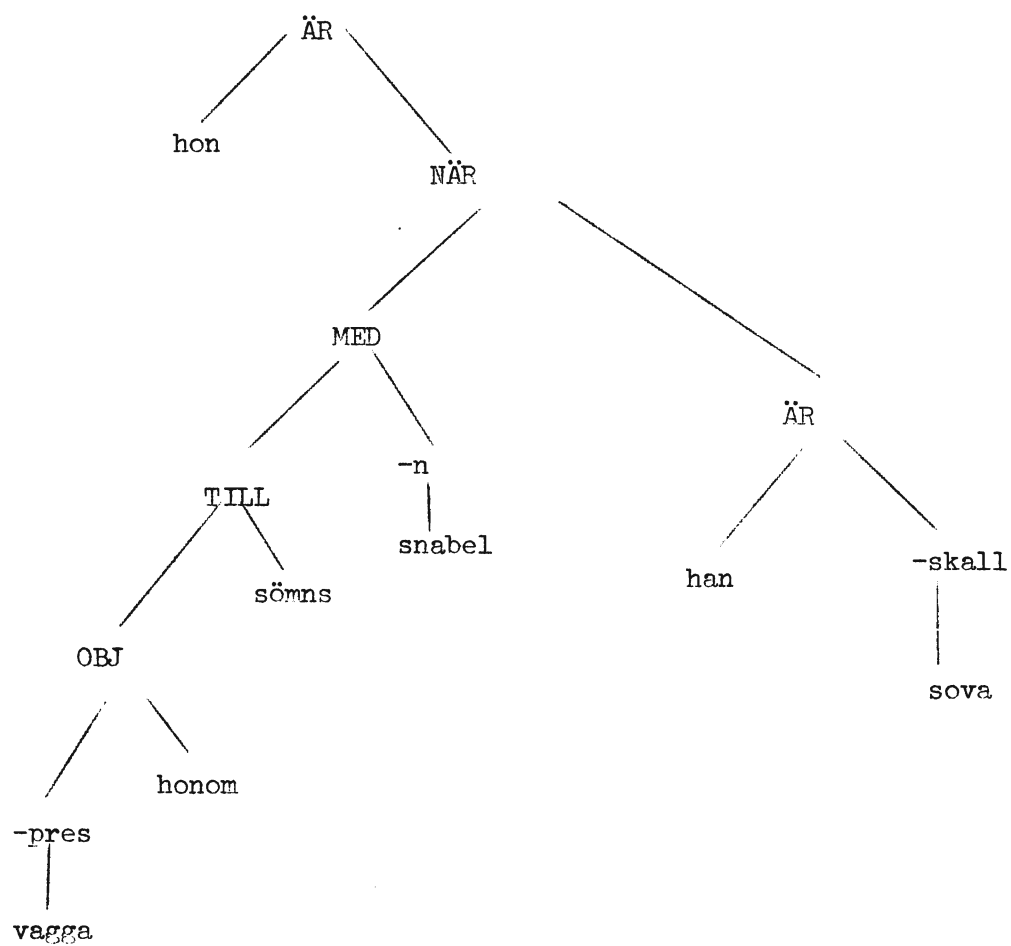
1. En liten BOS elefant ÄR bo -pres I
 Det (djup AV Den stor BOS skog)



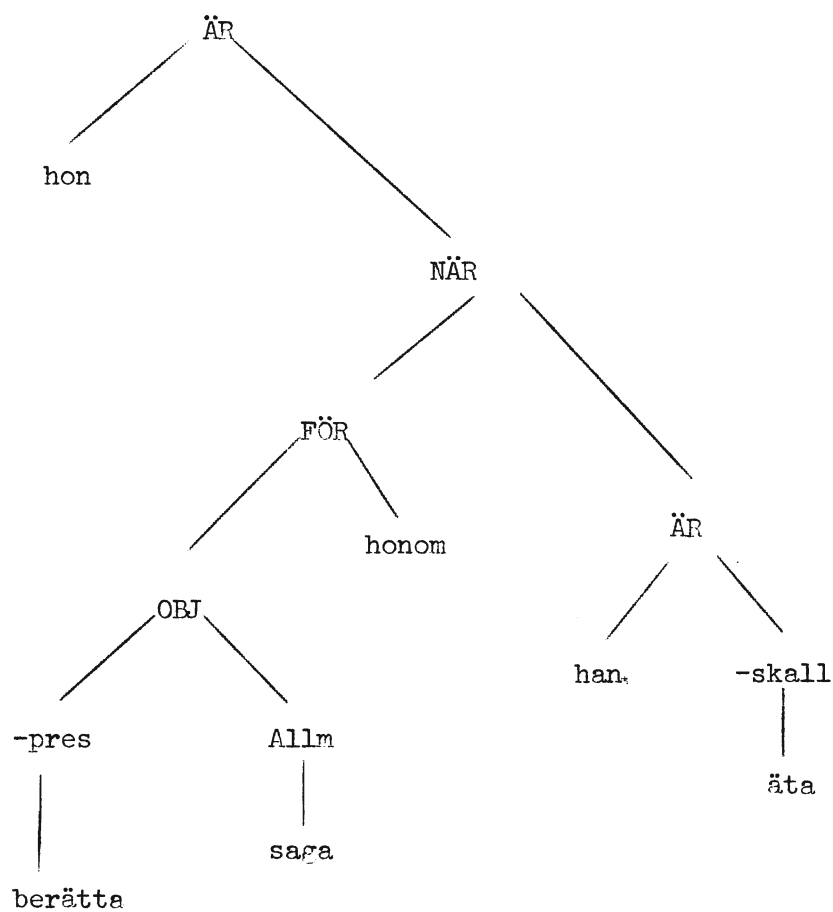
3. Hans mamma ÄR tycka-om -pres SÅ
 mycket OBJ Sin (liten BOS elefant)



4. hon ÄR vagga -pres OBJ honom TILL
 sömns MED snabel -n NÄR (han
 ÄR sova -skall)



5. hon ÄR berätta -pres OBJ Allm säga
FÖR honom NÄR (han ÄR äta -skall)



6. Uppgifter för senare steg i språkanalysen

Här ska jag bara räkna upp några olika saker som behöver göras, och ge en gissning av var de bäst kan göras.

1. Häng på ett predikatet "Santatt" på de uttryck som genereras. Görs som strukturtransformation.
2. Omvandla "Santatt m ÄR p" till "m ÄR p" (för lagring som binär relation i nätverk). Görs som strukturtransformation, ev. genererad från ett axiom.
3. Tyd BOS till OCK, BL, eller SS (dvs. fastställ hur adjektiv relaterar till varandra och till sitt substantiv). Görs bäst under acceptansen medelst slutsatsdragning.
4. Tyd -pres till antingen "brukas alltid" eller "görs just nu". Görs under acceptansen medelst slutsatsdragning.