STOCKHOLMS UNIVERSITET
Institutionen för Lingvistik

Victoria Johansson
VT-00

NP-detektion

UTVÄRDERING OCH FÖRSLAG TILL FÖRBÄTTRINGAR
AV GRANSKAS NP-REGLER

ABSTRACT

There were two aims of this essay. The first aim was to evaluate the performance of the rules for automatic detection of swedish noun phrases in *Granska*. The second aim was to provide suggestions to improvements of the existent rules. *Granska*, a project under development at NADA at KTH in Stockholm, is a program for grammar checking in an environment for writing support. It uses a mixed system with a probabilistic tagger and hand-coded linguistic rules for parsing at the phrase level. The rules for detection of noun phrases are an integrated part in the functions for grammar checking. The present rules detect ill-formed noun phrases but are under development so they will detect well-formed noun phrases as well. This is important because they can then be implemented as an aid in other rules for grammar checking and contribute to the improvement of the program. About 1000 noun phrases have been tagged manually and categorized in order to make an evaluation of the rules possible. Since Swedish noun phrases has complicated structures this categorization made it easier to see which types of NPs the program was succesful or not in detecting and also facilitated the production and improvements of the rules. In spite of the fact that the rules were under development when writing this essay both recall and precision was found to be about 80 percent. How to interpret the results is discussed more in detail in the essay.

PK i ASV med inriktning mot datorlingvistik
c-uppsats, 10 poäng
Handledare: Ola Knutsson

INNEHÅLLSFÖRTECKNING

1. INLEDNING....3
2. SYFTE....3
3. BAKGRUND....3
  3.1 Granska-projektet....4
  3.2 Granskas regelspråk....4
    3.2.1 Operatorer....5
    3.2.2 Vänsterledet: syntax och notation....5
    3.2.3 Sekvensvariabler....7
    3.2.4 Högerledet: syntax och notation....7
    3.2.5 Olika typer av regler med exempel....8
4. METOD....9
  4.1 Taggning....10
  4.2 Kategorisering....12
5. RESULTAT AV UTVÄRDERING....13
  5.1 Täckning....14
  5.2 Precision....15
  5.3 Kommentarer till utvärderingen....16
6. FÖRBÄTTRINGAR AV REGLER....18
  6.1 Apposition....18
  6.2 Egennamn i genitiv....19
  6.3 Egennamn med framförställda attribut....20
  6.4 Samordnade nominalfraser....21
  6.5 Frågande/relativa pronomen, determinerare och possesiva pronomen....23
  6.6 Adverb....24
7. SAMMANFATTNING....25
REFERENSER....27
Appendix 1: Exempel på kategoriserade NP....28
Appendix 2: Särdragsklasser och särdragsvärden i Granskas regelspråk....30
Appendix 3: Förbättrade och förändrade regler....33

1. INLEDNING

En NP-detektor är ett program som automatiskt detekterar nominalfraser. Eftersom nominalfrasen är en av den svenska grammatikens viktigaste byggstenar så är en väl fungerande NP-detektor ett bra hjälpmedel i grammatikkontrollprogram. Med hjälp av nominalfrasen kan man hitta andra språkliga enheter i texten, kongruensfel m.m. Men att automatiskt kunna detektera nominalfraser är användbart i flera sammanhang. NP-detektion är t.ex. ett viktigt verktyg vid informationsextrahering eftersom det är nominalfrasen som benämner det man refererar till. *"As the bulk of terminology consists of NP's- 80-90% depending on the source- fishing for NP's as a starting point for term extraction can be considered a justifiable approach."* (Arppe, 2000-10-07:1) Ett annat område där NP-detektion kan användas är vid översättarstöd där nominalfrasen, i stället för ordet, får fungera som översättningsenhet.

2. SYFTE

Det finns egentligen två syften med den här uppsatsen. Huvudsyftet är att göra en utvärdering av de regler för NP-detektion som Granska-projektet utvecklat. I detta ingår bl.a. att ta reda på vilka typer av nominalfraser NP-reglerna klarar av att detektera och vilka typer de inte klarar av samt om de övergenererar, d.v.s. ger några falska alarm. Det andra syftet är att ge förslag till utbyggnad och förbättringar av reglerna för NP-detektion. Här kommer jag också att visa på några av de problem som finns vid automatisk detektion av nominalfraser.

Eftersom NP-reglerna är under utveckling så kommer en del regler att ha hunnit förändras när den här uppsatsen är färdig. De regler jag utvärderar och förbättrar här är reglerna så som de såg ut i oktober 1999.

3. BAKGRUND

Att konstruera ett program för automatisk detektion av svenska nominalfraser är inte lätt då svenska nominalfrasers struktur är invecklad. Huvudordet kan bestå av ett substantiv, egennamn eller ett pronomen som kan ha såväl framförställda som efterställda attribut. De framförställda attributen kan delas in i sex huvudgrupper; totala, determinativa, possesiva, kvantitativa, selektiva och komparativa. (Teleman, 1969:5) De efterställda attributen kan bestå av prepositionsfraser, infinitivfraser, adjektiv-, adverb- eller participfraser och olika typer av bisatser. (Teleman, Hellberg, Andersson, 1999:86) En nominalfras kan också själv fungera som såväl framförställt som efterställt attribut till en annan nominalfras. Gunnel Källgren har beskrivit strukturen hos de svenska nominalfraserna och svårigheterna med att detektera dem;

> *"a highly interesting chaos. (…) Swedish noun phrases are a dream (or a nightmare) for any constructor of finite state machines. Number, gender, definitness and sometimes case in article, adjective and noun interact in complicated patterns, where a parsing process must often keep several different patterns active at the same time."* (Källgren, 1992:13)

The MorP parser (Källgren, 1992) utvecklades som ett projekt vid Stockholms Universitet och är ett exempel på en parser som nästan helt och hållet utgår ifrån ytkriterierna i språket. Detta gäller såväl den morfologiska som den syntaktiska informationen och den har ett mycket begränsat lexikon till hjälp. I denna parser ingår ett NP-program som klarar av att detektera ensamma

substantiv och substantiv med framförställda attribut med hög precision men inga NP med efterställda attribut.
En annan NP-detektor, som utger sig för att vara en av de bästa inom området, är Lingsofts *NPtool*. (Arppe, 2000-10-07) Den detekterar engelska nominalfraser och har informationsextrahering som huvudsyfte. Denna parser är så gott som helt regelstyrd. Den morfologiska analysen görs med hjälp av ett lexikon på 56000 engelska ordstammar men för okända ord används även heuristiska metoder[1] för att gissa rätt ordklass. Disambiguering och syntaktisk analys bygger på Constraint Grammar. *NPtool* klarar förutom framförställda attribut även av att detektera vissa prepositionsfraser som fungerar som efterställda attribut.

3.1 Granska-projektet

Den här uppsatsen handlar om de regler för NP-detektion som Granska-projektet vid NADA på KTH i Stockholm håller på och utvecklar. Granska är ett program för svensk datorstödd språkgranskning. Det ser ut som ett vanligt ordbehandlingsprogram men har dessutom en språkgranskningsfunktion som analyserar och markerar felaktigheter i texten om så önskas. Fel som programmet klarar av att detektera är t.ex. stavfel, felaktigt skrivna tecken, stilavvikelser och grammatikfel. Fel och problem i texten markeras med olika färgkoder beroende på vilken typ av fel det är. Information om det markerade området, kommentarer, rättningssförslag och länkar till källreferenser, t ex Svenska språknämndens skrivregler, ges också. Reglerna för NP-detektion ingår som en del av grammatikgranskningsfunktionen i Granska[2].

Granska är både probabilistiskt och regelstyrt, det har en probabilistisk tagger medan parsning på frasnivå är regelstyrd. För att programmet ska kunna hitta felaktigheter i texten krävs att varje ord i texten är försedd med ordklassinformation. Den probabilistiska taggare som Granska använder sig av är en andra ordningens Markovmodell som försöker hitta de mest sannolika taggarna för ambiguösa ord. Statistik förs på alla ordsekvenser på två och tre taggar och de sannolikaste taggarna beräknas sedan utifrån denna statistik. Mer än 95 procent av taggarna blir rätt med denna metod. Förutom ambiguösa ord finns många nya ord och sammansättningar som försvårar taggningen. Utöver den probabilistiska taggningen görs därför en morfologisk analys av ordet som även den bygger på statistik. Rätt ordklass kan på så sätt gissas med ganska stor precision. När taggningen väl är färdig detekteras felaktigheter i texten med hjälp av en uppsättning regler som beskriver olika typer av fel. Reglerna matchas sedan mot taggade ordsekvenser.

Inom Granska-projektet finns sedan länge regler som detekterar felaktiga nominalfraser, t.ex. kongruensfel som **den stora huset*. Nu arbetar man med att skriva om dessa regler så att de även detekterar korrekta nominalfraser. Detta är viktigt eftersom reglerna som detekterar nominalfraser då kan användas för att förbättra andra språkgranskningsregler i Granska.

3.2 Granskas regelspråk

Följande är inte ett försök att göra en fullständig beskrivning av Granskas regelspråk utan endast av det som är relevant för att förstå hur reglerna för NP-detektion fungerar. Reglerna i Granska är

[1] Att tagga en text heuristiskt eller probabilistiskt betyder att man använder sig av statistik för att räkna ut rätt ordklass.
[2] Det primära syftet med reglerna för NP-detektion i Granska är att fungera som en hjälp till andra språkgranskningsregler. Det går därför inte att jämföra dessa regler med en NP-detektor som har nominalfras-detektion som huvudsyfte.

inspirerade av objektorienterade programmeringsspråk. De består av ett vänsterled och ett högerled där vänsterledet beskriver vad som ska matchas i texten medan högerledet beskriver vad som skall göras med det som matchats. Det som ska matchas ses som ett objekt som kan beskrivas med hjälp av särdrag.

3.2.1 Operatorer

Det finns en rad operatorer i regelspråket. Vissa måste finnas med för att reglerna ska accepteras av programmet medan andra är optionella. Några av de här vanligt förekommande är följande:

& logiskt och mellan villkor
| logiskt eller mellan villkor
! logisk negation av villkor
= lika med
!= inte lika med
+ binärt plus
- binärt och unärt minus
:= tilldelningsoperator
, kommatecken avskiljer matchningsvariabler
; semikolon innebär logiskt eller mellan regler
--> en konsekvenspil som skiljer vänsterled från högerled
{ regelbörjan
} regelslut

3.2.2 Vänsterledet: syntax och notation

Följande element måste ingå i vänsterledet:

{ regelbörjan
X( ), matchningsvariabel, om det kommer flera avskiljs de med kommatecken
Y( )
--> konsekvenspil som avskiljer vänsterled från högerled

Matchningen i vänsterledet sker mot token[3]. Varje token ses som ett objekt och beskrivs med en matchningsvariabel, t.ex. X. Objektet har en rad attribut; textattribut och särdragsattribut. Genom att beskriva dessa attribut kan matchning av objektet göras. Själva villkoren för vilka attribut det matchade objektet ska ha sätts inom parentes efter variabelnamnet och flera attribut kan kombineras med hjälp av de logiska operatorerna. Särdragsattributen består av särdragsklasser som tilldelas olika särdragsvärden. En fullständig lista med dessa finns i appendix men här är några av dem:

Särdragsklasser	**Betydelse**	**Särdragsvärden**
gender	genus	utr, neu, utr/neu, mas
num	numerus	sin, plu, sin/plu
spec	species	ind, def, ind/def
case	kasus	nom, gen

[3] Token är de enheter i texten som märks upp med taggar, alla ord tilldelas ordklasstaggar men även punkter, kommatecken m.m. får sina speciella taggar.

wordcl	ordklass	nn, pm, jj rg, ro, pc, ab, dt, hd, ps, hs, pn, kn, pp

wordcl	**betydelse**
nn	nomen (substantiv)
pm	egennamn
jj	adjektiv
rg	räkneord: grundtal
ro	räkneord: ordningstal
pc	particip
ab	adverb
dt	determinerare
hd	frågande/relativ determinerare
hp	frågande/relativt pronomen
ps	possesiv
hs	frågande/relativ possesiv
pn	pronomen
kn	konjunktion

Parentesen efter matcningsvariabeln måste finnas med även om det inte finns några matchningsvillkor, det som matchas är då ett godtyckligt token:

X(wordcl=nn)   (matchar ett substantiv)
X( )   (matchar ett godtyckligt token)

Särdragsattributen har i följande exempel kombinerats med den logiska operatorn & (logiskt *och*), d.v.s. båda villkoren måste uppfyllas för att matchningen ska genomföras:

X(wordcl=dt & gender=utr),
Y(wordcl=nn & gender=utr)

Dessa variabler matchar alltså en determinerare i utrum samt ett substantiv i utrum. Ett sätt att jämföra värden på särdrag, om man t.ex. vill matcha en nominalfras där genus hos determinerare och substantiv kongruerar, är att använda punktnotation:

X(wordcl=dt),
Y(wordcl=nn & gender=X.gender)

Detta matchar endast korrekta nominalfraser som *en hus* och inga felaktiga som **ett hus* eftersom genus hos substantivet måste kongruera med determineraren. Punktnotation går bara att använda för redan introducerade variabler, det skulle alltså inte gå att vända på ordningen här, d.v.s. sätta punktnotationen i första variabelns villkor. Att jämföra värden hos särdrag går också bra att göra i högerledets notation (se nedan). Genom att använda operatorn != (inte lika med) går det även att uttrycka att attributet inte får ha ett visst värde:

X(wordcl!=jj)

Textattributen fungerar på liknande sätt som särdragsattributen. Om man vill matcha ordet *nominalfras* så skriver man detta ord inom citationstecken:

X(text=”nominalfras”)

3.2.3 Sekvensvariabler

Ibland vill man kunna beskriva flera token med samma variabel om matchningskraven för dessa token är desamma. Detta är möjligt att göra genom att sätta en operator efter parentesen och på detta sätt ange hur många token som ska matchas.

*	noll eller flera token
+	ett eller flera token
?	noll eller ett token
n	n står för ett heltal och sekvensen med token som matchas är noll till n lång

3.2.4 Högerledet: syntax och notation

Högerledet består av en rad fält. Endast ett av dessa är obligatoriskt, nämligen fältet action som anger vilken typ av regel det är. Högerledet börjar efter avskiljaren -->.

-->	avskiljare
action( )	i fältet action anges vilken typ av regel det är
}	regelslut

De olika (optionella) fälten i högerledet anger vad som ska göras med det som matchats i vänsterledet. När det gäller detektion av felaktiga nominalfraser kan man korrigera dem eller ge rättningsförslag m.m. i dessa fält. När det gäller detektion av korrekta nominalfraser däremot är det främsta syftet i det första skedet att detektera dem. Framförallt är det följande fält som är intressanta för detektion av korrekta nominalfraser:

corr( )	Det här är egentligen ett fält där ersättningsförslag kan anges då eventuella felaktigheter påträffats. Men det går naturligtvis bra att här ange de objekt man vill ska returneras utan att göra några förändringar. Funktionen concat som sätter ihop två argument till ett och returnerar detta som en textsträng kan också användas här om man t.ex. vill sätta hakparenteser runt nominalfrasen.
jump( )	Regelordningen är viktig i Granskas regler. I detta fält kan man manipulera med denna på flera sätt (se nedan).
info( )	Här kan man skriva kommentarer till det som matchats, t.ex. vilken typ av NP det är.
action( )	De regeltyper som är intressanta här är framförallt help (hjälpregler) och scrutinizing (grammatikkontrollregler).

Det är som sagt viktigt i vilken ordning man skriver reglerna. De exekveras i den ordning som de står i (top-down) vilket naturligtvis påverkar resultatet. När inga fler regler finns kvar att exekvera flyttas matchningen fram ett token om inget annat anges. Om man t.ex. har en regel som detekterar nominalfraser med framförställda attribut och en regel som detekterar nominalfraser utan framförställda attribut så kommer hela nominalfrasen med attribut att detekteras av den första regeln och sedan huvudordet ensamt av den andra regeln. I fältet jump kan man påverka ordningen på regelappliceringen på följande sätt:

1) jump(läge) Här anges läget för den regel som man vill ska appliceras härnäst på det/de token som matchats.
2) jump(endlabel) Detta betyder att inga fler regler ska appliceras på det som har matchats. Endlabel är regelsamlingens slut och regelexekveringen börjar om från regelsamlingens början och flyttar då automatiskt fram ett token i texten.
3) jump(läge, antal token) Här anges dels läget för den regel som ska appliceras härnäst och dels hur många token regelmatcharen ska hoppa framåt i texten.

3.2.5 Olika typer av regler med exempel

Det är framförallt regeltyperna scrutinizing och help som används i NP-reglerna. Scrutinizing är regler för stavnings- och grammatikkontroll och help är hjälpregler. Hjälpregler är regler som används i andra regler. Det här är en för enkelhetens skull förminskad variant av hjälpregeln *NPmin* (minimal nominalfras) i Granskas NP-regler:

```
NPmin@
{
X(wordcl=dt)?,
Y(wordcl=jj | wordcl=ro)*,
Z(wordcl=nn)
-->
info(”NP”)
action(help, case:=Z.case)
}
```

Denna hjälpregel matchar noll eller en determinerare följd av noll eller flera adjektiv eller räkneord samt ett substantiv. Eftersom det är en hjälpregel kommer denna regel ensam inte att uföra någon matchning om den inte används i en annan regel. Hjälpregeln sätts då inom parentes med variabelnamnet efter ett snedstreck men fungerar för övrigt ungefär på samma sätt som vilken annan matchningsvariabel som helst:

```
{
(NPmin/X)()
-->
corr(concat(concat(”[”, X.text), ”]NPmin”))
jump(endlabel, X.no_of_tokens)
action(scrutinizing)
}
```

Det som returneras blir här hela nominalfrasen, oavsett hur lång den är, eftersom variabeln X syftar till hela det matchade området i hjälpregeln *NPmin*. Det går också bra att kombinera hjälpregler med andra hjälpregler eller nya matchningsvillkor:

```
{
(NPmin/X)(case=gen),
(NPmin/Y)()
-->
corr(concat(”[”, X.text) concat(Y.text,”]NPnngen”))
```

```
jump(endlabel, X.no_of_tokens + Y.no_of_tokens)
action(scrutinizing)
}
```

Här har samma hjälpregel använts två gånger för att detektera nominalfraser som t.ex. *denna tidens hårda krav*, d.v.s. en NP i genitiv och en i nominativ. Det går som sagt att jämföra värden på särdrag även i högerledets notation. Här har man angett att kasus på den första nominalfrasen ska vara genitiv. Att detta lyckas beror på att det i hjälpregeln *NPmin* är angivet i fältet action (i högerledet) att kasus ska kongruera med huvudordets kasus.

En annan typ av regler som kan vara användbara är optionella hjälpregler. Den enda skillnaden i notationen jämfört med vanliga hjälpregler är att ett frågetecken sätts efter. Detta betyder att regeln kan förekomma men inte behöver göra det i frasen som ska matchas. I nominalfraser finns en rad attribut som kan förekomma men inte måste göra det, dessa skulle t.ex. kunna skrivas i en optionell hjälpregel.

Det finns ett par operatorer som kan användas mellan regler. Subtraktion uttrycks med tilde. A~B innebär att det som detekteras av regel A inte ska detekteras av regel B. Subtraktion anges mellan dubbla måsparenteser. Union mellan regler uttrycks med semikolon. A;B innebär att elementen måste detekteras antingen av regel A eller av regel B eller av båda reglerna (motsvarar logikens inklusiva *eller*) för att matchningen ska exekveras.

Huvudsyftet med Granskas NP-regler är att fungera som ett hjälpmedel i andra regler. Här är ett exempel på hur NP-regeln används i en regel som ska detektera och korrigera felaktig kongruens i predikativ eftersom man vill att den NP som kommer före kopulaverbet ska kongruera med adjektivet. (Domeij & Knutsson, 2000-01-26:2)

```
{
T(wordcl!=pp),
(NP),
(PP)+,
v3(wordcl=ab)?,
V(wordcl=vb)?,
X(wordcl=vb & vbt=kop),
Y(wordcl=jj & (gender!=NP.gender | num!=NP.num)),
Z(wordcl!=nn & wordcl!=jj)
-->
mark(all)
corr(T NP PP V3 V4 X if NP.spec=def then Y.get_form(gender:=NP.gender,
num:=NP.num, spec:=ind) else Y.get_form(gender:=NP.gender, num:=NP.num) end Z)
info("Substantivfrasen" NP.text "stämmer inte överenns med adjektivet" Y.text)
action(granskning)
}
```

4. METOD

För att kunna göra en utvärdering av NP-reglerna har det varit nödvändigt att märka upp nominalfraser för hand i ett antal texter för att ha något att testa dem emot. Det är två saker man måste ta ställning till innan man sätter ihop en textkorpus; för det första hur stor den ska vara och för det andra vilken typ av texter den ska innehålla.

Hur många nominalfraser räcker det att märka upp för att få fram tillförlitliga siffror på täckning och precision? Eftersom det är ett ganska tidskrävande arbete att märka upp nominalfraser för hand och uppsatsens omfång är begränsat bestämde jag mig för att märka upp 1000 nominalfraser till att börja med. Texternas omfång var då på sammanlagt 2824 ord vilket kanske verkar lite om man jämför med Lingsofts *NPtool* som använt sig av texter på 20000 ord i sin utvärdering. Det visade sig dock att det räckte med de totalt 1077 uppmärkta nominalfraserna för att få en ganska bra bild av NP-reglernas utförande. Det finns olika metoder för att mäta täckning och precision. Metoden jag använder mig av har jag hämtat från Lingsoft då det verkade vara en överskådlig och lätt metod och dessutom vanlig i dessa sammanhang. Metoden beskrivs närmare i utvärderingen.

Även när det gällde ställningstagandet om vilket material korpusen skulle innehålla var jag tvungen att ta hänsyn till uppsatsens begränsade storlek. Att försöka täcka in alla genrer för att få en balanserad korpus hade lett till att den blivit alldeles för stor. De texter som använts här är uteslutande tidningstexter från Dagens Nyheter och Forskning & Framsteg. Trots detta begränsade urval är korpusen inte heller specialiserad i strikt mening då texterna består av såväl nyheter som reportage. De uppmärkta nominalfraserna kan ändå antas vara typiska för dessa genrer och skulle säkert varit något annorlunda om textkorpusen bestått av t.ex. skönlitterära texter.

Förutom taggningen har nominalfraserna delats in i olika kategorier. För det första fungerade kategoriseringen som en hjälp i utvärderingen, för att se vilka kategorier NP-reglerna klarar respektive inte klarar att detektera. Kategoriseringen var också till stor hjälp vid regelskrivandet eftersom det gav en bra bild av strukturen hos nominalfraserna.

4.1 Taggning

På grund av nominalfrasernas struktur är det inte självklart hur man ska märka upp dem. Ett uttryck man ibland stöter på i dessa sammanhang är ”maximal length NP’s”. En maximalt lång NP kan bestå av ett antal ingående nominalfraser. Frågan är då om man ska märka upp den längsta nominalfrasen, de ingående nominalfraserna eller både och. Beroende på hur man väljer att göra blir antalet uppmärkta fraser olika och detta påverkar i slutändan resultatet av utvärderingem. Om NP-reglerna t.ex. detekterar de längsta nominalfraserna men inte de ingående och man vid taggningen märkt upp även de ingående nominalfraserna så kommer resultatet av utvärderingen att bli sämre än om man hade märkt upp endast de längsta fraserna. Därför anser jag det vara viktigt att beskriva hur själva taggningen gått till.

En nominalfras är uppbyggd kring ett huvudord som ensamt kan utgöra hela nominalfrasen men som också kan ha framförställda och/eller efterställda attribut. De framförställda attributen har fått "hänga ihop" med huvudordet så att hela frasen taggas som en NP. *Den nuvarande ledningen* bildar alltså endast en NP även om *ledningen* självt också är en nominalfras. Likadant har jag gjort i de fall då en nominalfras i genitiv fungerar som attribut till en annan nominalfras som i *kommunförbundets färska rapport*.

När huvudordet även har efterställda attribut har först hela nominalfrasen med alla framförställda och efterställda attribut taggats som en NP. När det efterställda attributet består av en apposition, som i t.ex. *FN-majoren Jörgen Öberg* har ingen vidare uppdelning gjorts utan hela frasen har fått bilda en NP. När de efterställda attributen däremot består av predikativa attribut, bisatser eller satsförkortningar så har även de ingående nominalfraserna i de eftersällda attributen taggats som egna NP. I den relativa bisatsen i nominalfrasen *de artiklar som innehåller det svåra ordet* taggas även *det svåra ordet* som en NP så att hela frasen totalt taggas som tre nominalfraser. NP med adverbattribut fanns endast en i textkorpusen, *50 felstavningar till*, och där bildar inte attributet någon nominalfras.

När det gäller prepositionsattributen så har taggningen blivit lite olika beroende på hur många prepositionsattribut som funnits i frasen. Om nominalfrasen endast har en prepositionsfras som attribut, som t.ex. i *en videoupptagning av programmet*, så taggas hela frasen plus nominalfraserna *en videoupptagning* och *programmet* så att det totalt blir tre NP. I nominalfrasen *tillverkarnas förklaring till bristen på de nya mobiltelefonmodellerna* däremot så har både *tillverkarnas förklaring* och *bristen* prepositionsattribut. Här har inte *bristen på de nya mobiltelefonmodellerna* taggats som en egen NP. De som taggas som nominalfraser är hela frasen samt *tillverkarnas förklaring*, *bristen* och *de nya mobiltelefonmodellerna*, d v s sammanlagt fyra.

I samordnade nominalfraser taggas både de ingående nominalfraserna och hela frasen så att *kvalitet och mångfald* bildar tre nominalfraser. Ibland kan nominalfrasen sakna huvudord. Det är t.ex. fallet i frasen *en av de kidnappade* där *en* bildar en (elliptisk) NP.

Vad gäller pronomenet *som* så har det inte taggats som en NP. Det är visserligen ett pronomen men refererar inte till någonting och det är ändå det som är nominalfrasernas funktion, att beteckna det man refererar till.

Jag gjorde en grov indelning av de flesta av nominalfraserna redan vid taggningen. De fyra kategorierna jag delade in dem i då var; NPdef (definita), NPind (indefinita), NPpn (pronomen) och NPpm (egennamn).

Exempel på taggning:

```
NPdef[Glädjen] stod högt i NPind[tak] på NPpm[Wall Street] när
NP[NPdef[inflationssiffrorna] för NPdef[september]] offentliggjordes.

NP[NPdef[Initiativtagaren] och NPdef[förlagschefen Jan-Erik Pettersson]]
motiverar NPdef[urvalet], däribland NP[NPpm["Persiska antologin"] av NPpm[Eric
Hermelin]] och NP[NPpm["Samhället som teater"] av NP[NPpm[Ingemar Karlsson] och
NPpm[Arne Ruth]]], med att NPpn[de] är tidlösa och har NP[NPind[en aktualitet]
som kan väcka NPind[intresse] hos NPind[en bred läsekrets]].

Däremot vet NPpn[vi] inte NP[NPpn[vilka] som uppskattar respektive irriteras av
NPind[grafiskt splittrade texter]].

På NPdef[torsdagen] publicerades NP[NPind[en mycket stor undersökning] i
NP[NPpm[New England Journal of Medicine], NPind[en av NPdef[världens främsta
medicinska tidskrifter]]]].

NP[NP[NPdef[Informationschefen] på NPpm[Swedint] i NPpm[Södertälje]] NPpm[Claes
Wolgast]] bekräftar för NPpm[TT] att NP[NPind[en] av NPdef[de kidnappade]] är
svensk men NPpn[han] vill inte avslöja NPdef[mannens identitet] innan
NPind[anhöriga] informerats.

NP[NPind[Vilken vinkel] NPdef[fotografen] väljer] kan prägla NP[NPdef[hela
bilden] av NPind[en händelse]].
```

```
Utan NPpn[detta] kommer NPpn[ingen] att kunna dömas för NPdef[mordet].
```

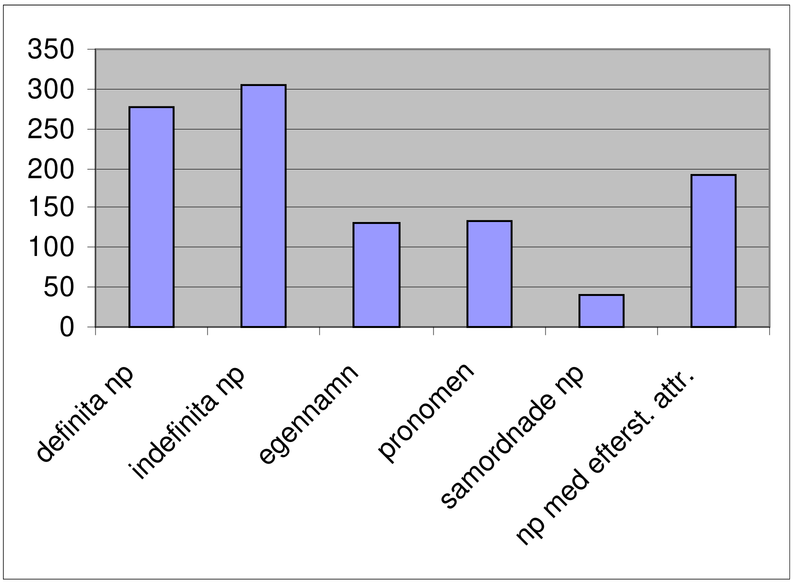
4.2 Kategorisering

När alla texter var taggade delades nominalfraserna in i ytterligare kategorier. Inspirationen till indelningen har jag framförallt fått från Svenska Akademiens grammatik, men då det främsta syftet med kategoriseringen var att underlätta utvärderingen och regelskrivandet och inte att göra en ”korrekt” analys av nominalfraser så överensstämmer inte alltid denna indelning med Svenska Akademiens indelning. (Teleman et al., 1999) Egennamn och personliga pronomen utan attribut skall väl egentligen räknas till definita NP eftersom referenten då oftast är känd men jag har ändå valt att ha dem i egna kategorier. Observera att detta endast gäller ensamma namn och pronomen. De saknar oftast attribut men i de fall de haft sådana har även dessa delats in i definita och indefinita NP. Interrogativa nominalfraser, t.ex. *vilket medieföretag*, fanns inte så många och de har placerats bland de indefinita nominalfraserna utan kvantitetsattribut istället för i en egen grupp. Tabellen nedan visar vilka kategorier de uppmärkta nominalfraserna är indelade i med exempel och antal.

1. DEFINITA NP		**278**
1:1 Enkel med definithetsattribut	den rätta skärpan, landets kommuner	(90)
1:2 Enkel utan definithetsattribut	balanskravet, nästa år	(179)
1:3 Komplex	alla nyhetsprogrammen	(3)
1:4 Elliptisk	den genomarbetade	(1)
1:5 Med egennamn som huvudord	Erikssons T28	(4)
1:6 Utan subst., egennamn eller substantiviskt pron. som huvudord	det här	(1)
2. INDEFINITA NP		**304**
2:1 Enkel med kvantitetsattribut	en aktualitet, fem titlar	(119)
2:2 Enkel utan kvantitetsattribut	kommentarer, verksamma författare	(176)
2:3 Elliptisk	sex (av FN:s militärobservatörer)	(7)
2:4 Med egennamn som huvudord	brun Saab 9000	(2)
3. EGENNAMN	Champions League, Wolgast	**130**
4. PRONOMEN	han, det, varandra	**133**
5. SAMORDNADE NP	Socialstyrelsen och Livsmedelsverket	**40**
6. NP MED EFTERSTÄLLDA ATTRIBUT		**192**
6:1 Apposition	byn Azhara, b-vitaminet folsyra	(32)
6:2 Prepositionsattribut	definitionen av “pocketboken”	(100)
6:3 Relativa bisatser	platsen där gisslan hålls	(49)
6:4 Infinitivattribut	vanliga sätt att beskriva grupper	(3)
6:5 Adverbattribut	50 felstavningar till	(1)
6:6 Predikativa attribut	En 30-årig man, tidigare känd av polisen	(1)
6:7 Satsförkortning	det han antyder	(6)

Några nominalfraser var svårare att kategorisera än andra. En av dem var *det här med millennieskiftet*. Skulle *det här* placeras bland pronomenen eller bland de elliptiska nominalfraserna? I Svenska Akademiens grammatik, under rubriken "Definita nominalfraser utan substantiv, egennamn eller substantiviskt pronomen som huvudord", finns exempel på fraser konstruerade med infinitivfras eller narrativ bisats som apposition; *"Om nominalfrasens huvudord är ett definit pronomen i neutrum singularis (vanligen det här, det där) inleds i ledigt språk infinitivfrasen eller den narrativa satsen ofta av med; det här med att starta redan i år".* (Teleman et al., 1999:42) Jag placerade pronomenet i ovan nämnda kategori men satte hela frasen bland nominalfraserna med prepositionsattribut. En annan svårplacerad nominalfras var *en röd eller brun Saab 9000.* Man kanske kan betrakta den som en enkel indefinit nominalfras med kvantitetsattribut, men eftersom den innehåller en konjunktion ligger det nära till hands att betrakta den som en samordnad nominalfras. Om man betraktar den som en samordnad nominalfras så måste man kategorisera även de samordnade fraserna *en röd* och *brun Saab 9000*. Jag valde det senare så att *en röd* fick bilda en elliptisk indefinit NP och *brun Saab 9000* en indefinit NP med egennamn som huvudord. Det sista exemplet på en svårdefinierad NP är titeln *Samhället som teater*. Den blev taggad som egennamn. Några av de uppmärkta nominalfraserna passade in i flera kategorier. Ett exempel är *en mycket stor undersökning i New England Journal of Medicine, en av världens främsta medicinska tidskrifter*. Här finns både prepositionsattribut och apposition i samma NP.

Av det totala antalet nominalfraser i korpusen är ungefär 80 procent nominalfraser utan eller med framförställda attribut. Bland nominalfraser med efterställda attribut har mer än hälften prepositionsattribut. Följande tabell visar hur antalet nominalfraser fördelar sig i grupperna definita (cirka 26 procent), indefinita (cirka 28 procent), egennamn (cirka 12 procent), pronomen (cirka 12 procent), samordnade NP (cirka 4 procent) och NP med efterställda attribut (cirka 18 procent).



Tabell 1: Antal uppmärkta NP i respektive kategori.

5. RESULTAT AV UTVÄRDERING

Metoden jag har använt mig av för att räkna ut täckning och precision är hämtad från Lingsofts utvärdering av sin *NPtool*. (Voutilainen, 2000-01-06:1) Definitionen är följande;

Täckning: förhållandet ”antalet NP som detekteras / totala antalet NP”.
Precision: förhållandet ”antalet NP som detekteras / totala antalet detekterade NP”.

Det innebär att man för att få fram täckning dividerar summan av alla korrekt detekterade nominalfraser med summan av det totala antalet uppmärkta nominalfraser. För att få fram precision divideras istället denna summa med summan av den totala mängden detekterade nominalfraser. Med korrekt detekterad nominalfras menas här de NP som överensstämmer med de i korpusen uppmärkta nominalfraserna. Detta bör man ha i åtanke vid utvärderingen eftersom såväl täckning som precision hade kunnat gynnas eller missgynnas av en annan indelning av nominalfraserna än den som gjorts här. Antalet NP som inte räknas med i mängden ”antalet NP som detekteras”, d.v.s. de som inte överensstämmer med de på förhand uppmärkta nominalfraserna, är följande:

- NP som inte detekteras alls[4].
- NP som är felaktigt/delvis detekterade, d.v.s. om till exempel endast huvudordet detekteras men inte de framförställda attributen. Då är den detekterade nominalfrasen egentligen en korrekt NP men trots detta räknas den alltså som felaktig eftersom jag har märkt upp alla NP med framförställda attribut som endast en NP.
- Annat än NP som detekteras.

Om en NP både detekteras korrekt och felaktigt så räknas den felaktiga detektionen till "det totala antalet detekterade NP". I denna summa finns alltså mängden övergenereringar som görs. Även sådant som inte är nominalfraser men som ändå detekteras är naturligtvis övergenereringar. Efter utförd granskning med NP-reglerna på de aktuella texterna har jag kunnat dra nedanstående slutsatser om Granskas regler för NP-detektion vad gäller täckning och precision.

5.1 Täckning

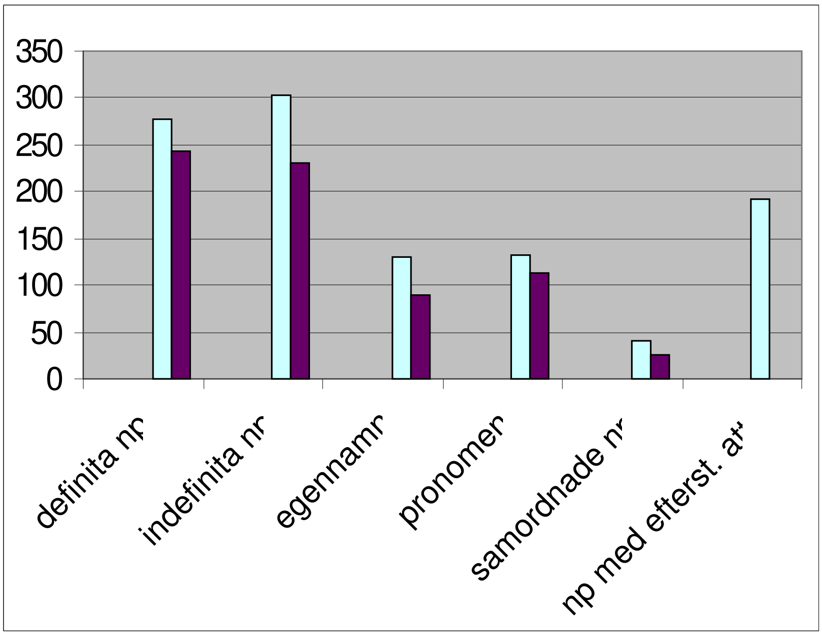
Av de totalt 1077 för hand uppmärkta nominalfraserna detekterar Granskas NP-regler 704 av dessa. Då är alla nominalfraser medräknade, alltså de utan attribut, de med framförställda attribut och de med efterställda attribut. Efterställda attribut är ju de som är svårast att hantera när det gäller detektion på automatisk väg. Granskas NP-regler klarar inte av att detektera några av dessa vilket helt enkelt beror på att det inte finns några regler implementerade ännu för att detektera dem. Med ett undantag dock, det finns en regel som ska detektera vissa typer av NP med prepositionsattribut (t.ex. *en av de politiska tänkarna*) men den fungerar inte särskilt bra. Ungefär tio NP detekteras av denna regel, samtliga felaktiga[5].

```
{
X(wordcl=dt | wordcl=pn),
Y(wordcl=pp),
(Npmin/Z)()
-->
```

Totalt har jag märkt upp 192 nominalfraser med efterställda attribut. Det är alltså inte så stor del av nominalfraserna som har efterställda attribut. Majoriteten av dessa har prepositionsattribut. Här måste man komma ihåg att när det gäller just prepositionsattributen så taggades bara den längsta nominalfrasen. Hade även de av de ingående nominalfraserna som hade prepositionsattribut fått bilda egna NP så hade den totala mängden nominalfraser med efterställda attribut blivit större. Eftersom NP-reglerna inte detekterar några av dessa heller så hade följaktligen täckningen också

[4] Att nominalfraser inte detekteras alls beror ibland på reglerna men kan också bero på felaktig taggning av texten.
[5] Det fanns förslag till förbättringar av denna regel som inte var implementerade ännu när denna utvärdering gjordes.

blivit något sämre. Mängden är dock så pass liten att skillnaden hade varit marginell. Om man däremot räknar bort alla NP med efterställda attribut och bara ser till täckningen när det gäller att detektera nominalfraser utan attribut eller med framförställda attribut så blir resultatet avsevärt bättre. Av totalt 885 nominalfraser detekteras då 704 av dessa korrekt, d.v.s. ungefär 80 procent. Om man tittar på hur NP-reglerna klarar att detektera de olika kategorierna med NP var för sig så finner man det allra bästa resultatet bland de definita nominalfraserna där 244 av 278 (cirka 88 procent) blir korrekt detekterade. Efter de definita nominalfraserna klarar NP-reglerna övriga grupper i följande ordning; pronomen 113 av 133 (cirka 85 procent), indefinita NP 231 av 304 (cirka 76 procent), egennamn 90 av 130 (cirka 69 procent), samordnade NP 26 av 40 (65 procent). Tabell 2 visar det totala antalet uppmärkta NP i första kolumnen och det totala antalet korrekt detekterade NP i den andra kolumnen för respektive kategori.



Tabell 2: Antal detekterade NP i respektive kategori.

5.2 Precision

Den totala mängden detekterade NP efter utförd granskning var 847. Genom att dela antalet korrekt detekterade NP (d.v.s. 704) med denna summa fick jag fram att precisionen ligger på ungefär 80 procent. Av de nominalfraser som inte räknas in i mängden "antalet NP som detekteras" ovan (373 stycken) är de allra flesta delvis/felaktigt detekterade. Det är ganska få nominalfraser som inte detekteras alls. En delvis detekterad nominalfras är t.ex. *arbete* där den korrekta nominalfrasen är *mycket arbete*. Att den totala mängden detekterade NP är större än antalet korrekt detekterade NP betyder att vissa övergenereringar gjorts. Ett exempel på övergenerering är när frasen blir både korrekt och felaktigt detekterad som *Christer Petterssons uttalande* där förutom den korrekta detektionen även *Petterssons uttalande* detekteras som en NP. Det totala antalet detekterade NP blir med andra ord fler. Det kan också vara så att nominalfrasen delas upp i fler delar än jag hade tänkt mig. Framförallt gäller detta vid apposition, t.ex. *kulturminister Margot Wallström*, där ungefär hälften detekteras som två NP istället för som en (med huvudordet för sig och appositionen för sig).

Några fall med detektion av sådant som ej är nominalfraser finns också och dessa detekteras nästan uteslutande av den regel som ska detektera NP med vissa typer av prepositionsattribut. Ett exempel på detta är **den av honom*. Annars beror de flesta av dessa fall på felaktig taggning av den detekterade frasen och inte på reglerna för NP-detektion.

5.3 Kommentarer till utvärderingen

När det gäller nominalfraserna med framförställda attribut och de med apposition så beror majoriteten av de ej detekterade nominalfraserna inte på att de inte detekteras alls utan på att de detekteras felaktigt. Några exempel på sådant som ställer till problem är nominalfraser i genitiv, adverb och nominalfraser med flera substantiv eller egennamn.

NP-reglerna klarar av att detektera nominalfraser med en NP i genitiv som attribut men inte då detta attribut i sin tur också har en NP i genitiv som attribut. Så här ser matchningskravet i regeln som detekterar nominalfraser som har nominalfraser i genitiv som framförställda attribut ut;

(NPmin/X)(case=gen),
(NPmin/Y)()

Nominalfrasen *Sveriges televisions planer* detekteras därför som *televisions planer* eftersom det i hjälpregeln *NPmin* endast kan/måste förekomma ett substantiv, egennamn eller pronomen. Anledningen till detta är att man inte vill att fraser som t.ex. **hon den* i *fick hon den* ska detekteras. Frågan är hur många NP i genitiv som kan förekomma före huvudordet. Eftersom det är en hjälpregel som detekterar dem kan man inte använda operatorn * (noll eller flera) efter matchningsvillkoret utan endast ? (noll eller en) vilket inte skulle hjälpa i det här fallet.

I hjälpregeln *NPmin* finns inga adverb bland matchningsvillkoren varför *den mycket större grupp kvinnor* detekteras som **större grupp*. Här det fullständiga matchningsvillkoret i hjälpregeln *NPmin*:

```
NPmin@
{
X(wordcl=dt | wordcl=ps | wordcl=hd)?,
Y(wordcl=jj | wordcl=ro | wordcl=rg | wordcl=pc)*,
Z(wordcl=nn | wordcl=pm | wordcl=pn)
```

Däremot finns såväl räkneord som adjektiv med bland attributen i regeln varför det är svårare att förklara att den detekterar *det 100 meter långa borrhålet* som *det*, *100 meter* och *borrhålet*. Här borde determineraren detekteras tillsammans med *100 meter* och *långa* som är ett adjektiv tillsammans med *borrhålet*. Troligtvis är determineraren (*det*) taggad som pronomen eftersom det inte har samma genus som det närmaste substantivet (*meter*).

Att flera substantiv eller egennamn följer på varandra är ganska vanligt. Vi har redan sett två exempel på detta, *grupp kvinnor* och *meter(…)borrhålet*. Det här förekommer också i nominalfraser med apposition, t.ex. *själva ordet millennium*, och i definita och indefinita NP med egennamn som huvudord, t.ex. *en nedgången Christer Pettersson*.

Vad gäller grupp 1:2, definita NP utan definita attribut, grupp 2:2, indefinita NP utan kvantitetsattribut samt egennamn så är orsaken till att nominalfraser inte detekteras i många fall att de detekteras av regeln som detekterar samordnade nominalfraser. Nominalfrasen *propositionen och tillståndsvillkoren* detekteras som en NP varför de två samordnade nominalfraserna *propositionen* och *tillståndsvillkoren* inte detekteras var för sig som egna NP. Regeln som detekterar samordnade NP står högre upp i regelsamlingen och i fältet jump är det angett att inga fler regler ska appliceras på de token som matchats.

Fallet kan också vara det omvända, nämligen att nominalfraser detekteras av flera regler. Den samordnade nominalfrasen *Ingemar Karlsson och Arne Ruth* detekteras dels (felaktigt) av regeln som detekterar samordnade nominalfraser som *Karlsson och Arne* och de båda egennamnen detekteras (korrekt) av regeln som detekterar egennamn. Anledningen till att bara den ena hälften av egennamnen detekteras av regeln i det första fallet är återigen att endast ett substantiv, egennamn eller pronomen kan ingå i *NPmin* som fungerar som hjälpregel här:

```
NPkonj@
{
(NPmin/X)(),
Y(wordcl=kn & (text="och" | text="eller" | text="samt")),
(NPmin/Z)()
-->
```

En brist i regelspråket är att inte enbart den längsta matchningen detekteras. Framförallt ställer detta till problem vid detektion av egennamn. För att undvika att egennamn övergenereras finns för nuvarande två regler där den första tar ut NP med minst två egennamn och den andra NP med ett eller flera:

```
{
X(wordcl=pm)+,
Y(wordcl=pm)
-->
```

```
{
X(wordcl=pm),
Y(wordcl=pm)*
-->
```

En text har under rubriken författarnamnet, *Jorunn Amcoff*, och börjar med egennamnet *Bokförlaget Ordfront*. Eftersom det inte finns något skiljetecken emellan blir de detekterade nominalfraserna; *Jorunn Amcoff, Jorunn Amcoff Bokförlaget*, och *Jorunn Amcoff Bokförlaget Ordfront*. Det förekommer också att enbart en del av namnet detekteras. Ett exempel är nominalfrasen *Georgiens president Eduard Sjevardnadze* där *Eduard* inte detekteras utan endast *Sjevardnadze*.

När det gäller de självständiga pronomenen detekteras inte de relativa/interrogativa alls. Det beror på att de har en annan tagg än andra pronomen och denna tagg finns inte med i regeln som detekterar självständiga pronomen. Elliptiska NP detekteras inte heller eftersom det då varken finns något substantiv, pronomen eller egennamn i frasen vilket måste ingå för att de ska matchas av *NPmin*.

6. FÖRBÄTTRINGAR AV REGLER

Nedan följer en beskrivning av hur jag gått till väga för att förbättra några av reglerna för NP-detektion. Då syftet här inte bara är att visa vilka förbättringar som gjorts utan även att visa på vilka svårigheter som finns med NP-detektion så ges också visst utrymme till "diskussion" kring detta.

6.1 Apposition

Nominalfrasen *biträdande spaningsledaren Lars Jonsson* detekteras inte korrekt eftersom den består av ett substantiv och två egennamn. Det går inte att lösa detta genom att sätta operatorn + (en eller flera) i matchningsvillkoret så att fler än ett substantiv och/eller egennamn blir detekterade:

Z(wordcl=nn | wordcl=pm | wordcl=pn)+

Det leder till övergenereringar som *biträdande spaningsledaren* och *biträdande spaningsledaren Lars*. Orsaken är att det räcker med att ett substantiv eller egennamn hittas för att matchningskravet ska uppfyllas och frasen detekteras som NP enligt regeln. Det går inte heller att lösa genom att dela upp matchningsvillkoret i två matchningsvillkor med operatorn * (noll eller flera) efter det första:

Z1(wordcl=nn | wordcl=pm)*,
Z2(wordcl=nn | wordcl=pm | wordcl=pn)

Till skillnad från andra villkor som kommer före huvudordet så är matchningsvillkoret här detsamma för Z1 som för Z2 varför det räcker med att villkoret för Z2 uppfylls för att frasen skall bli detekterad. Detta leder alltså till samma övergenereringar som i föregående fall. Den enda lösningen som kvarstår om man vill undvika dessa övergenereringar är då att få med egennamnen på något annat sätt. Det finns för nuvarande två regler som detekterar egennamn. Att det behövs två regler beror på att man vill att de namn som består av två eller flera token ska detekteras före de med ett token så att *Claes Wolgast* detekteras av den första regeln och *Aktuellt* av den andra. Om det inte fanns två regler skulle de egennamn som består av fler än ett token övergenereras. Genom att göra om reglerna som detekterar egennamn till hjälpregler, *NPpm1* och *NPpm2*, kan de användas i andra regler:

```
NPpm1@
{
X(wordcl=pm)+,
Y(wordcl=pm)
-->
action(help, case:=Y.case)
}

NPpm2@
{
X(wordcl=pm),
Y(wordcl=pm)*
-->
```

```
action(help, case:=Y.case)
}
```

Nu är det möjligt att göra en regel som sätter ihop nominalfrasen med substantiv som huvudord med nominalfrasen med egennamn som huvudord till en NP. I själva verket måste det bli två regler även här eftersom det finns två regler för egennamnen:

```
{
(NPmin/X)(),
(NPpm1/Y)()
-->
corr(concat("[", X.text) concat(Y.text, "]NPapp"))
jump(endlabel, X.no_of_tokens + Y.no_of_tokens)
action(scrutinizing)
}

{
(NPmin/X)(),
(NPpm2/Y)()
-->
corr(concat("[", X.text) concat(Y.text, "]NPapp"))
jump(endlabel, X.no_of_tokens + Y.no_of_tokens)
action(scrutinizing)
}
```

Denna förändring gör nu att *biträdande spaingsledaren Lars Jonsson* detekteras som en NP istället för som två.

Att få till substantiven som apposition är betydligt svårare än egennamnen. Man kan göra en regel som ser likadan ut som den som tar ut nominalfraser med nominalfraser i genitiv som attribut fast utan villkoret att kasus ska vara i genitiv. Då blir nominalfraser som *b-vitaminet folsyra* t.ex. korrekt detekterade men även en del fraser som ej är nominalfraser som **det han* i frasen *det han antyder* eller **skatten nästa år* i *18 kommuner höjer skatten nästa år*. Möjligen skulle man kunna lösa detta genom att ha krav på att huvudordet måste vara i obestämd form. Dessvärre är just species mycket besvärligt eftersom det handlar så mycket om kontextuell information. I exemplet ovan är det attributet och inte bestämdheten hos huvudordet som gör att *nästa år* är en definit nominalfras.

6.2 Egennamn i genitiv

Genom att ta bort egennamnet (pm) som villkor i *NPmin* kan man åstadkomma flera förbättringar men vissa nya problem uppstår också. För det första kan man uppnå att fraser som *Christer Petterssons uttalanden* detekteras som en fras istället för som två. Den detekteras redan korrekt men dessutom felaktigt som *Petterssons uttalande*. Orsaken är att frasen detekteras av två regler. Först av en regel som har *NPmin* i genitiv som hjälpregel och sedan (korrekt) av en regel som har *PMgen* (detekterar noll till tre egennamn och ett egennamn i genitiv) som hjälpregel. Genom att stryka egennamnet ur *NPmin* blir frasen endast korrekt detekterad men ett nytt problem uppstår, nämligen att egennamn med framförställda attribut inte längre kommer att detekteras.

6.3 Egennamn med framförställda attribut

För att egennamn med framförställda attribut skall bli detekterade när man strukit attributet pm ur *NPmin* måste man göra nya hjälpregler och regler liknande de för *NPmin* men med två eller flera respektive ett eller flera egennamn som huvudord istället. Jag kallar de nya hjälpreglerna *NPmin2* och *NPmin3*. Förutom hjälpreglerna krävs även två nya regler där hjälpreglerna kan sättas in.

```
NPmin2@
{
X(wordcl=dt | wordcl=ps | wordcl=hd)?,
Y(wordcl=jj | wordcl=ro | wordcl=rg | wordcl=pc)*,
(NPpm1/Z)()
-->
action(help, case:=Z.case)
}

NPmin3@
{
X(wordcl=dt | wordcl=ps | wordcl=hd)?,
Y(wordcl=jj | wordcl=ro | wordcl=rg | wordcl=pc)*,
(NPpm2)()
-->
action(help, case:=Z.case)
}

{
(NPmin2/X)(case=nom)
-->
corr(concat(concat("[", X.text), "]NPmin2"))
jump(endlabel, X.no_of_tokens)
action(scrutinizing)
}

{
(NPmin3)(case=nom)
-->
corr(concat(concat("[", X.text), "]NPmin3))
jump(endlabel, X.no_of_tokens)
action(scrutinizing)
}
```

NPmin2 och *NPmin3* matchar nu nominalfraser med två eller flera respektive med ett eller flera egennamn som huvudord med eller utan framförställda attribut. Jag placerar dem före reglerna som detekterar egennamn så att inte *Tyskland* detekteras först och sedan *hela Tyskland* (vilket också händer när värdet pm ingår *NPmin* som kommer efter reglerna för egennamn).

När det gäller egennamn som består av fler än två namn så kommer dessa fortfarande att bli övergenererade om man inte skriver ytterligare en regel för tre namn eller flera som sätts före de andra i regelsamlingen. Detta är dock ganska ovanligt.

6.4 Samordnade nominalfraser

En annan sak man borde kunna komma till rätta med genom att stryka attributet pm är egennamnen i samordnade nominalfraser. Fraser som *Aktuellt och Rapport* fungerar bra nu men inte *Dagens Nyheter och Svenska Dagbladet* som detekteras **Nyheter och Svenska*. Man löser detta genom att skriva en ny regel för samordnade nominalfraser som detekterar två egennamn-konjunktion-två egennamn. Men om man stryker attributet pm ur *NPmin* måste man också skriva en ny regel som detekterar de två samordnade ensamma egennamnen. Och då kommer ändå inte fraser som *Andreas Carlgren och jag* bli detekterade.

Nu kommer hjälpreglerna *NPmin2* och *NPmin3* till användning igen eftersom jag kan använda dem även för att detektera de samordnade nominalfraserna. För att minska ned något på antalet regler gör jag två hjälpregler som ser ut på följande sätt;

```
NPhjälp1@
{
(NPmin)()-->action(help);
(NPmin2)()-->action(help)
}

NPhjälp2@
{
(NPmin)()-->action(help);
(NPmin3)()-->action(help)
}
```

Semikolon mellan reglerna innebär att den ena eller båda måste matchas för att regeln skall gälla. Här innebär det alltså att *NPhjälp1* matchar både nominalfraser med substantiv, pronomen och två eller flera egennamn som huvudord medan *NPhjälp2* matchar nominalfraser med substantiv, pronomen och ett eller flera egennamn som huvudord. Det är alltså egennamnen, att man vill detektera de längsta först, som gör att det måste bli två regler. Den befintliga hjälpregeln som matchar samordnade nominalfraser gör jag sedan om till två regler, *NPkonj1* och *NPkonj2*, och även den regel som genomför själva detektionen måste göras om till två regler;

```
NPkonj1@
{
(NPhjälp1/X)(),
Y(wordcl=kn & (text="och" | text="eller" | text="samt")),
(NPhjälp1/Z)()
-->
action(help)
}

NPkonj2@
{
(NPhjälp2/X)(),
Y(wordcl=kn & (text="och" | text="eller" | text="samt")),
(NPhjälp2/Z)()
-->
action(help)
}
```

```
{
(NPkonj1/X)()
-->
corr(concat(concat(" [", X.text), "]NPkonj"))
jump(endlabel, X.no_of_tokens)
action(scrutinizing)
}

{
(NPkonj2/X)()
-->
corr(concat(concat(" [", X.text), "]NPkonj"))
jump(endlabel, X.no_of_tokens)
action(scrutinizing)
}
```

Med dessa nya regler blir nu *Rapport och Aktuellt*, *Andreas Carlgren och jag* samt *Dagens Nyheter och Svenska Dagbladet* korrekt detekterade utan övergenereringar. Men fortfarande blir inte samordnade nominalfraser med uppräkningar, *rubriker, ingresser och allehanda illustrationer*, eller med genitivattribut, *en hel världs datorer och inbäddade mikrochip*, korrekt detekterade. Detta försökte jag lösa genom att göra en ny hjälpregel, *NPuppräkn*, som jag satte in i hjälpreglerna *NPhjälp1* och *NPhjälp2*:

```
NPuppräkn@
{
X(wordcl=pm | wordcl=nn |cht=mid)+
-->
action(help)
}

NPhjälp1@
{
(NPuppräkn)()-->action(help);
(NPmin)()-->action(help);
(NPmin2)()-->action(help)
}

NPhjälp2@
{
(NPuppräkn)()-->action(help);
(NPmin)()-->action(help);
(NPmin3)()-->action(help)
}
```

Nu blev uppräkningar visserligen detekterade men också nya övergenereringar av egennamn eftersom det nu kom in ett nytt pm-attribut i de båda NPhjälp-reglerna.

Ytterligare ett exempel på en samordnad NP som inte blir detekterad är *kön, ålder, intressen och andra vanliga sätt att beskriva grupper*. För att kunna detektera alla typer av samordnade nominalfraser måste man först ha regler som detekterar alla typer av nominalfraser som finns eftersom så gott som alla NP kan samordnas. Däremot gick det bra att få *Forskning & Framsteg* detekterad som en fras genom att lägga till textkravet & i hjälpreglerna *NPkonj1* och *NPkonj2*:

```
NPkonj1@
{
(NPhjälp1/X)(),
Y(wordcl=kn & (text="och" | text="eller" | text="samt" | text="&")),
(NPhjälp1/Z)()
-->
action(help)
}
```

6.5 Frågande/relativa pronomen, determinerare och possesiva pronomen

Interrogativa fraser som *vilket medieföretag* detekteras som *medieföretag* och självständiga frågande/relativa pronomen detekteras inte alls. De frågande/relativa pronomenen *vilken* och *vilket* taggas ibland som hp och ibland som hd. Attributet hd finns redan i *NPmin* men jag lägger nu till även attributen hp och hs (frågande/relativ possesiv). Jag måste lägga till textkravet "text!="som"" eftersom *som* taggas som hp. Jag tar också bort attributet pn ur *NPmin* eftersom pronomen inte har framförställda attribut. Efter att jag har tagit bort särdragsvärdena pm och pn och lagt till de nya värdena ser hjälpregeln nu ut så här:

```
NPmin@
{
X(wordcl=dt | wordcl=ps | wordcl=hd | wordcl=hs | wordcl=hp & (text!="som"))?,
Y(wordcl=jj | wordcl=ro | wordcl=rg | wordcl=pc)*,
Z(wordcl=nn)
-->
action(help, case:=Z.case)
}
```

De nya värdena bör man nog också lägga till i *NPmin2* och *NPmin3*. Även om det inte finns några sådana exempel i denna textkorpus så är det helt möjligt att konstruera nominalfraser med frågande/relativa pronomen/possesiv eller possesiv och egennamn.

I regeln som detekterar pronomen lägger jag till attributen hd och hp samt samma textvillkor som jag lade till i *NPmin*.

```
{
X(wordcl=pn | wordcl=hd | wordcl=hp & (text!="som"))
-->
corr(concat("[", X.text) concat(Y.text, "]NPint"))
jump(endlabel, X.no_of_tokens + Y.no_of_tokens)
action(scrutinizing)
}
```

Jag byter också plats på reglerna så att *NPmin* exekveras innan regeln för pronomen för att på så vis slippa övergenereringar.

6.6 Adverb

Nominalfraser som *den mycket större grupp kvinnor* och *betydligt färre barn* blir som sagt felaktigt detekterade p.g.a. adverben. Att bara lägga till ett adverb bland attributen i NPmin-reglerna borde vara lätt men så är inte fallet. Anledningen är att adverben ibland bestämmer verbet och ibland ett adjektiv i nominalfrasen. Att bara lägga till attributet ab i *NPmin* leder därför till att fraser som **däribland Persiska Antologin* och **ungefär 40 barn* detekteras. Man skulle här behöva göra någon form av kontextkänsliga regler så att bara adverb efter determinerare eller bara adverb framför adjektiv detekteras. Inga av de nuvarande reglerna i NP-detektorn är i strikt mening kontextkänsliga. (Internt inom nominalfrasen kan man säga att de är kontextkänsliga men det finns i de nuvarande reglerna inga villkor för vad som kan eller inte kan förekomma utanför nominalfrasen.) Det finns vissa möjligheter att skriva kontextkänsliga regler som ser till vad som finns utanför frasen som ska detekteras. Den kontext man är intresserad av här ingår ju i frasen som ska detekteras varför det inte går att tillämpa dessa här. Det man också skulle kunna göra är att skriva ytterligare regler där antingen determineraren eller adjektivet ingår som obligatoriskt element istället för som optionellt. Det skulle bli ganska många regler eftersom det redan finns tre NPmin-hjälpregler och det skulle behövas tre nya. Dessutom skulle alla regler där någon av NPmin-reglerna ingår behöva skrivas om. Det allra bästa här vore istället att skriva en hjälpregel, *AP*, som detekterar adjektivfraser och använda denna som optionell hjälpregel i hjälpreglerna för nominalfraser. Även de andra attributen kan man skriva om till en hjälpregel, *Attr*, och sedan sätta in bägge dessa hjälpregler i de tre hjälpreglerna *NPmin*, *NPmin2* och *NPmin3*:

```
Attr@
{
X(wordcl=dt | wordcl=ps | wordcl=hd | wordcl=hs | wordcl=hp & (text!="som"))?,
Y(wordcl=ro | wordcl=rg | wordcl=pc)*
-->
action(help)
}

AP@
{
X(wordcl=ab)*,
Y(wordcl=jj)
-->
action(help)
}

NPmin@
{
(Attr/X)()?,
(AP/Y)()?;
Z(wordcl=nn)
-->
action(help)
}

NPmin2@
{
(Attr/X)()?,
(NPpm1/Y)()
-->
```

```
action(help, case:=Z.case)
}

NPmin3@
{
(Attr/X)()?,
(NPpm2/Y)()
-->
action(help, case:=Z.case)
}
```

Nu detekteras bara de adverb som bestämmer adjektiv, som t.ex. i *en mycket stor undersökning*, och inga andra.

7. SAMMANFATTNING

Syftet med uppsatsen var att göra en utvärderng av de regler för NP-detektion som Granska-projektet på NADA på KTH håller på och utvecklar för att ha som hjälp i sina språkgranskningsregler. Jag har för detta ändamål märkt upp 1077 nominalfraser för hand i en textkorpus bestående av cirka 2824 ord samt kategoriserat nominalfraserna i olika kategorier. Reglerna har sedan testats på textkorpusen för att se vilka typer av nominalfraser de klarar av respektive vilka typer de inte klarar av att detektera. Vidare har jag gett förslag till förbättringar av de befintliga reglerna för NP-detektion och diskuterat vilka svårigheter som finns med detta.

Att detektera nominalfraser på automatisk väg är inte lätt då de svenska nominalfraserna har en ganska invecklad struktur. Av de för hand uppmärkta nominalfraserna i textkorpusen fanns åtminstone 20 olika typer. Då har ändå inte så mycket arbete lagts ned på själva klassificeringen eftersom detta inte var huvudsyftet med uppsatsen utan endast skulle fungera som en hjälp i utvärderingen och regelskrivandet. Huvudindelningen av nominalfraser brukar göras mellan definita och indefinita. Om man tittar närmare på dessa grupper ser man att det egentligen är ganska lite som utmärker dem. Både definita och indefinita NP kan ha såväl substantiv, pronomen och egennamn som huvudord men kan också sakna huvudord alldeles (ellips). Om huvudordet är ett substantiv är det oftast i bestämd form om nominalfrasen är definit men behöver inte vara det och samma sak gäller för de indefinita nominalfraserna. Vad gäller attributen så kan dessa vara såväl framförställda som efterställda och se ut på en mängd olika sätt beroende på bestämdhet, numerus, kasus o.s.v. De efterställda attributen hör till de svårare när det gäller automatisk detektion av nominalfraser och det är få NP-detektorer i dag som klarar av att detektera några av dessa.

Granskas NP-regler klarar inte av att detektera några nominalfraser med efterställda attribut eftersom det inte finns några sådana regler implementerade. Bland nominalfraserna utan attribut eller med framförställda attribut så detekterar reglerna definita nominalfraser bäst (cirka 88 procent) och samordnade nominalfraser sämst (cirka 65 procent). Att resultatet blev sämst på de samordnade nominalfraserna är förklarligt eftersom de flesta typer av NP kan samordnas, alltså även NP med efterställda attribut. Om man tittar på det sammanlagda resultatet av reglernas utförande så ligger såväl täckning som precision på ungefär 80 procent då man räknat bort nominalfraser med efterställda attribut. Detta tycker jag måste anses vara ett ganska bra resultat med tanke på att NP-reglerna är långt ifrån färdigutvecklade. Många regler var under utveckling men ej implementerade ännu i den version som här testats och utvärderats.

Att göra en utvärdering är inte helt lätt då resultatet så mycket beror på hur man märkt upp nominalfraserna från början. En annan uppmärkning än den jag gjort här hade kanske gett ett något annorlunda resultat. Jag har därför inte bara presenterat resultatet av utvärderingen i siffror utan även försökt visa hur detektionen faktiskt gått till, vilka regler som detekterar vad och varför. Hur man ska märka upp nominalfraser tror jag handlar mycket om vilket syfte som finns med att detektera dem. Om syftet t.ex. är informationsextraktion är det troligt att man vill kunna söka på alla de i nominalfrasen ingående nominalfraserna. Här är syftet grammatikkontroll, att hitta felaktigheter inom nominalfrasen, varför det borde vara av större intresse att få hela nominalfrasen detekterad som en NP.

Jag har också gett förslag till förbättringar av reglerna för NP-detektion. Det är många saker man måste tänka på vid regelskrivandet. Vilken ordning man skriver reglerna i är mycket viktigt. Man måste hela tiden hålla reda på vad de andra reglerna detekterar så att reglerna komplementerar varandra och inga övergenereringar görs. Detta är lättare sagt än gjort eftersom det fort blir ganska många regler. Några av de saker jag lyckats förbättra är; detektion av egennamn som apposition, detektion av fler än ett egennamn i samordnade nominalfraser, detektion av ensamma eller framförställda frågande/relativa pronomen eller determinerare och detektion av adjektivfraser som framförställt attribut.

En stor brist i Granskas regelspråk anser jag är att inte endast den längsta matchningen detekteras. Detta får till följd att onödigt många regler måste skrivas om man vill undvika övergenereringar. Ett bra exempel är detektion av egennamn där man inte vet hur många egennamn frasen består av. För nuvarande måste man skriva olika regler där den första i ordningen tar ut de NP som har flest egennamn o.s.v. Detta påverkar också alla de regler som detekterar nominalfraser där egennamn kan ingå och det är ganska många.

En bra hjälp i detektionen av nominalfraser skulle vara om man även skrev regler som detekterar andra typer av fraser. Jag har i mina förslag till förbättringar skrivit en hjälpregel som tar ut adjektivfraser för att på detta sätt få med adverben bland de framförställda attributen. Även prepositionsfraser ingår ju i många nominalfraser och en regel som tar ut dessa skulle kunna användas som hjälpregel i reglerna för NP-detektion. Att använda sig av de möjligheter som finns att skriva kontextkänsliga regler skulle troligen också kunna förbättra utförandet. Varken de regler som utvärderats eller de regler jag själv skrivit är kontextkänsliga.

REFERENSER

Arppe, A. 1999-10-07. Term Extraction from Unrestricted Text. [WWW document]. URL http://www.lingsoft.fi/doc/nptool/term-extraction.html, NODALIDA-95

Domeij, R. & Knutsson, O. 2000-01-26. Granska-ett effektivt hybridsystem för svensk grammatikkontroll. [WWW document] URL http://www.nada.se/theory/projects/granska/rapporter/nodalidaabstract.html, Nada,KTH

Kann, V. 1999-12-14. Populär beskrivning av Granska. [WWW document]. URL http://www.nada.kth.se/theory/projects/granska/popular.html

Knutsson, O. 1999. Granskas regelspråk, tentativ version, 1999-10-14. IPLab, Nada, KTH

Källgren, G. 1992. Making maximal use of surface criteria in large-scale parsing: the MorP parser. PILUS 60, Institute of Linguistics, Stockholm University.

Teleman, U. 1969. Definita och indefinita attribut i nusvenskan. Studentlitteratur, Lund

Teleman, U., Hellberg, S., Andersson, E. 1999. Svenska Akademiens grammatik, del 3: Fraser. Svenska Akademien, Stockholm

Voutilainen, A. 2000-01-03. Overview. [WWW document] URL http://www.lingsoft.fi/doc/nptool/intro/overview.html

Voutilainen, A. 2000-01-06. Performance. [WWW document] URL http://www.lingsoft.fi/doc/nptool/intro/performance.html

Appendix 1: Exempel på kategoriserade NP

1. DEFINITA (med substantiv som huvudord)

1:1 Enkel med definithetsattribut
Christer Petterssons uttalande
FN:s militärobservatörer
1990-talets tidningsboom
sina uppgifter
den rätta skärpan
den person
denna information

1:2 Enkel utan definithetsattribut
torsdagskvällen
samma källa
Europabörserna
tiotiden
90-talet
informationsmängden
desktop publishingtekniken
kväll

1:3 Komplex med pronominella attribut
alla nyhetsprogrammen
hela bilden

DEFINITA (utan substantiv som huvudord)

1:4 Elliptiska
den genomarbetade

1:5 Med egennamn som huvudord
nya SVT24
den instängde Sune Scott
Nokias 3210

1:6 Utan substantiv, egennamn eller pronomen som huvudord
det här

2. INDEFINITA (med substantiv som huvudord)

2:1 Enkel med kvantitetsattribut
inget bevisvärde
alla politiska partier
tiotusentals offentliganställda
8-10 beväpnade män
allt fler grafiska styrsignaler
fyra gånger
många kvinnor
ett par år
en massiv informationskampanj

2:2 Enkel utan kvantitetsattribut
minnesluckor
viktiga händelser
olika ställen
full tid
kinesiska forskare
läsvanor
ny studie
konkreta bevis

INDEFINITA (utan substantiv som huvudord)

2:3 Elliptisk

en (eller två nyhetsredaktioner)
sex (kommer att sänka skatten)
78 (av landets kommuner)

2:4 med egennamn som huvudord

en nedgången Christer Pettersson

3. EGENNAMN

TV3
Stockholm
TT
Globen
Stockholms universitet
Abchazien
Forskning & Framsteg

4. PRONOMEN

det
han
något
oss
detta(2)
alla
ingen
vem

5. SAMORDNADE NOMINALFRASER

nyheter och samhällsfrågor
Abchazien och omvärlden
sex av FN:s militärobservatörer och en tolk

6. EFTERSTÄLLDA ATTRIBUT

6:1 Apposition

biträdande spaningsledaren Lars Jonsson
siffran 2 procent
sökmotorn AltaVista

6:2 Prepositionsattribut

en helt ovetenskaplig penetrering av landets större databaserade textarkiv
en frist till klockan tolv idag svensk tid
plats för kommentarer av andra idag verksamma författare och specialister inom respektive område
extra intag av folsyra

6:3 Relativa bisatser

vem som mördade Olof Palme
det riksdagsbeslut som styr tillståndsvillkoren
de södra delarna där förekomsten är densamma som i Sverige

6:4 Infinitivattribut

uppgiften att garantera kvalitet och mångfald på den svenska tv-marknaden
hans chanser att återväljas idag

6:5 Adverbattribut

50 felstavningar till

6:6 Predikativt attribut

En 30-årig man, tidigare känd av polisen

6:7 Satsförkortning

den information tittaren får
det han sagt

Appendix 2: Särdragsklasser och särdragsvärden i Granskas regelspråk

särdragsklass	**betydelse**	**särdragsvärde**
gender	genus	utr, neu, utr/neu, mas
num	numerus	sin, plu, sin/plu
spec	species (bestämdhet)	ind, def, ind/def
case	kasus	nom, gen
vbf	verbform	prs, prt, inf, sup, imp
mood	modus	kon
pef	perfekt form	prf
voice	verbalgenus (diates)	akt, sfo
deg	komparationsgrad	pos, kom, suv
pnf	pronomenform	sub, obj, sub/obj
wordcl	ordklass	nn, pm, jj, rg, ro, vb, pc, ab, in, ha, dt, hd, ps, hs, pn, hp, sn, kn, pp, ie, dl, pl, uo, an?
vbt	verbtyp	aux, kop
cht	teckentyp	mad, mid, pad
sed	meningsavgränsare	sen
style	stiltyp	datm, foal, frmo, fsms, lgpp, libb, lprs, onfl, pavb, psvb, stbb, svba, vard
phrase	frastyp	npdef, npindef
cap	gemener/versaler	nocapped, firstcapped, allcapped, mixcapped
rgt	räkneordstyp	yea
nntype	substantivtyp	set, dat

wordcl

särdragsvärde	**betydelse**	**exempel**
nn	nomen (substantiv)	bil
pm	egennamn	Lars
jj	adjektiv	grön
rg	räkneord, grundtal	12
ro	räkneord, ordningstal	första
vb	verb	springa
ab	adverb	mycket
in	interjektion	ja
ha	frågande/relativt adverb	när
dt	determinerare (artikel)	den
hd	frågande/relativ determinerare	vilken
ps	possesiv	hennes
hs	frågande/relativ possesiv	vems

pn	pronomen	hon
hp	frågande/relativ pronomen	vem
sn	subjunktion	om
kn	konjunktion	och
pp	preposition	till
dl	skiljetecken	.
pc	particip	springande
pl	partikel	om
uo	utländskt ord	the
an	förkortning	d. v. s.

gender

utr	utrum	en
neu	neutrum	ett
utr/neu	utrum/neutrum	de
mas	maskulinum	bleke

num

sin	singular	bil
plu	plural	bilar
sin/plu	singular/plural	hans

spec

ind	indefinit	en
def	definit	den
ind/def	indefinit/definit	gula

case

nom	nominativ	hunden
gen	genitiv	hundens

nntype

set	måttapposition	antal
dat	datumord	lördag

vbf

prs	presens	spelar
prt	preteritum	spelade
inf	infinitiv	spela
sup	supinum	spelat
imp	imperativ	spela

mood

kon	konjunktiv	vare

pef

prf	perfekt	kastad

voice

akt	aktiv	spela
sfo	s-form	spelas

deg

pos	positiv	vacker
kom	komparativ	vackrare
suv	superlativ	vackrast

pnf

sub	subjektsform	jag
obj	objektsform	mig
sub/obj	subjekt/objekt	den

vbt

aux	hjälpverb	**har** druckit
kop	kopula	bollen **är** grön
mod	modalt	**kan** dricka

rgt

yea	årtal	1954

cht

mad	skiljetecken vid meningsslut	.!?
mid	skiljetecken inom en mening	,
pad	diverse tecken	”[]()’

sed

sen	meningsbörjan och meningsslut	

Appendix 3: Förbättrade och förändrade regler

```
{
(NPkonj1/X)()
-->
corr(concat(concat(" [", X.text), "]NPkonj"))
jump(endlabel, X.no_of_tokens)
action(scrutinizing)
}

{
(NPkonj2/X)()
-->
corr(concat(concat(" [", X.text), "]NPkonj"))
jump(endlabel, X.no_of_tokens)
action(scrutinizing)
}

{
(NPmin/X)(),
(NPpm1/Y)()
-->
corr(concat("[", X.text) concat(Y.text, "]NPapp"))
jump(endlabel, X.no_of_tokens + Y.no_of_tokens)
action(scrutinizing)
}

{
(NPmin/X)(),
(NPpm2/Y)()
-->
corr(concat("[", X.text) concat(Y.text, "]NPapp"))
jump(endlabel, X.no_of_tokens + Y.no_of_tokens)
action(scrutinizing)
}

{
(NPmin2/X)(case=nom)
-->
corr(concat(concat("[", X.text), "]NPmin"))
jump(endlabel, X.no_of_tokens)
action(scrutinizing)
}

{
(NPmin3/X)(case=nom)
-->
corr(concat(concat("[", X.text), "]NPmin"))
jump(endlabel, X.no_of_tokens)
action(scrutinizing)
}

{
(NPpm1/X)()
-->
corr(concat(concat("[", X.text), "]NPpm1"))
```

```
jump(endlabel, X.no_of_tokens)
action(scrutinizing)
}

{
(NPpm2/X)()
-->
corr(concat(concat("[", X.text), "]NPpm2"))
jump(endlabel, X.no_of_tokens)
action(scrutinizing)
}

{
X(wordcl=pn | wordcl=hd | wordcl=hp &(text!="som"))
-->
corr(concat(concat("[", X.text), "]NPpn "))
jump(endlabel, X.no_of_tokens)
action(scrutinizing)
}

NPkonj1@
{
(NPhjälp1/X)(),
Y(wordcl=kn & (text="och" | text="eller" | text="samt" | text="&")),
(NPhjälp1/Z)()
-->
action(help)
}

NPkonj2@
{
(NPhjälp2/X)(),
Y(wordcl=kn & (text="och" | text="eller" | text="samt")),
(NPhjälp2/Z)()
-->
action(help)
}

NPmin@
{
(Attr/X)()?,
(APmin/Y)()?,
Z(wordcl=nn | wordcl=pn)
-->
action(help, case:=Z.case)
}

NPmin2@
{
(Attr/X)()?,
(NPpm1/Y)()
-->
action(help, case:=Z.case)
}
```

```
NPmin3@
{
(Attr/X)()?,
(NPpm2/Y)()
-->
action(help, case:=Z.case)
}

NPpm1@
{
X(wordcl=pm)+,
Y(wordcl=pm)
-->
action(help, case:=Y.case)
}

NPpm2@
{
X(wordcl=pm),
Y(wordcl=pm)*
-->
action(help, case:=Y.case)
}

NPhjälp1@
{
(NPmin)()-->action(help);
(NPmin2)()-->action(help)
}

NPhjälp2@
{
(NPmin)()-->action(help);
(NPmin3)()-->action(help)
}

Attr@
{
X(wordcl=dt | wordcl=ps | wordcl=hd | wordcl=hs | wordcl=hp & (text!="som"))?,
Y(wordcl=ro | wordcl=rg | wordcl=pc)*
-->
action(help)
}

APmin@
{
X(wordcl=ab)*,
Y(wordcl=jj)
-->
action(help)
}
```