GÖTEBORGS UNIVERSITET
INST FÖR SVENSKA SPRÅKET

GU-ISS-2012-03

Adverbialkarakteristik för praktisk informationsextraktion i svensk text
Projektrapport

Kenneth Wilhelmsson

Forskningsrapporter från institutionen för svenska språket, Göteborgs universitet
Research Reports from the Department of Swedish

ISSN 1401-5919

www.svenska.gu.se/publikationer/GU-ISS

Sammandrag

Den aktuella rapporten beskriver ett projekt som i första hand har inneburit ett praktiskt arbete syftande till att skapa en automatiserad process som returnerar frågeled, t.ex. *varifrån*, för adverbialled, t.ex. *inifrån rummet*, i svensk digital text. Det är en utbytesprocess som behövs av rent praktiska skäl i uppgiften frågegenerering, vilken innebär att en samling frågor som en text besvarar genereras snabbt automatiskt. Denna process finner sin plats i program som på olika sätt syftar till att ge informationsåtkomst i godtycklig okänd svensk text. Det är i detta tillämpningsfall fråga om att på något sätt öppna upp för den stora informationsmängd som i datalogiskt perspektiv ligger *'ostrukturerad'*, dvs. i naturligt språk-form.

Syftet med att avgöra lämpliga frågeled (ofta till en *hv*-form) för förekommande satsled i text har dock förmodligen en mer allmän relevans än användning i nämnda programtyp. Förutom att också behövas i andra liknande datalingvistiska applikationer kan själva frågeställningen rymmas inom ramarna för grundforskningen. De vanliga semantiskt grundade adverbialkategorierna (vilka skiljer sig åt mellan olika grammatikor) definierar gärna adverbialkategorier just genom att beskriva vilka slags frågor de besvarar. Att som här sikta på att avgöra frågeled för adverbial är en mer detaljerad uppgift än att avgöra adverbialkategori.

Den praktiska metod som implementerats i projektet kan sönderdelas i ett antal steg som antas vara allmängiltiga och svåra att undgå med det aktuella syftet. Indata till programmet är ett i princip godtyckligt adverbialled som användaren i prototypprogrammet kan skriva in. De nämnda steg som tar vid är de följande. 1) En uppmärkning med ordklass- och annan grammatisk information för varje löpord inleder. Detta sker med en statistisk trigrambaserad s.k. Hidden Markov-modell. 2/3) Ett avgörande av vilken strukturtyp som ledet har (bisats, PP, etc.) görs utifrån löporden med informationen i föregående steg. Intimt förknippat med denna uppgift är bestämning av huvudord, och för flera led även bestämning av andra signifikanta komponenter som rektionshuvudord. Lösningen till detta delsteg heter *rangbaserad chunkning*. 4) De steg som följer härefter skiljer sig mycket åt beroende på den aktuella strukturtypen. För prepositionsfraser undersöks t.ex. preposition och, beroende på vilken preposition det är fråga om, rektionshuvudord, dess grundform och andra ingående textsegment. I arbetet har t.ex. *SweFN* (Borin, Dannélls, Forsberg, Toporowska Gronostaj, & Kokkinakis, 2010) delvis undersökts för att eventuellt förbättra avgörandet av substantivsemantik, vilket ofta blir relevant för PP-adverbial.

Rapporten visar hur uppgiften praktiskt sett varierar mycket i svårighetsgrad, från de fall där adverbialet utgörs av t.ex. particip-, adverbfraser eller bisatser, då en mappning till motsvarande frågeled ofta kan ske direkt utifrån huvudordet – till de mest komplicerade fallen av PP och s.k. *som*-fraser där kombinationer av huvudord, rektionshuvudord, dess grundform samt annan syntaktisk och semantisk information krävs för att urskilja förekomsters särskilda frågemotsvarigheter. Ett återkommande tema i det praktiska arbetet är *undantag* som behöver kännas igen. Exempelvis kategorin satsadverbial, som kan anta många olika strukturella former men som ändå oftast renderar resultatet ’ingen frågemotsvarighet’, måste kännas igen uttryckligen (ev. tillsammans med andra med samma frågeledsresultat). Även processen som helhet bygger emellertid programmeringstekniskt på grundfall och undantag. I många fall, som t.ex. för *i*-PP finns det en mängd olika motsvarigheter och vad som får utgöra grundfall i programmet blir en empirisk/heuristisk fråga under det att regler skrivs mot faktiska förekomster av adverbial i Stockholm Umeå Corpus (Hädanefter *SUC*). Att *i* liksom andra prepositioner kan sägas ha en prototypisk riktningsbetydelse betyder inte att *var* nödvändigtvis ska fungera som utgångsfall. Det förekommer ’lager’ av undantag inom olika strukturslag i programmet men även externt motiverade sådana utgående från huvudverbet, som genom valensmatchning kan klargöra att ett adverbial är ’prepositionsobjekt’ och därmed får andra omfrågningsegenskaper. De användargränssnitt som skapats och använts för regelskrivande utifrån faktiska exempel har tillåtit viss omedelbar regeluppdatering och återkontroll vid åsynen av felaktiga resultat. Det är också genom tillägg av nya undantagsregler i någon mening som programmet rimligen ska kunna förbättras framöver från den aktuella kvalitetsnivån. Korrektheten som uppnåtts hittills är inte kvantitativt övertygande men detta arbete som saknar föregångare möjliggör kontinuerlig förbättring genom programmet.

Projektet visar att mappningsuppgiften i stora stycken verkar görbar när rätt identifikation av huvudord, rektionshuvudord etc. identifieras med hjälp av metoden ovan. Emellertid finns fall då det aktuella totala perspektivet, ”ge frågeled för samtliga adverbial”, gör att uppgiften känns märklig och då det är oklart vad som egentligen är korrekt frågemotsvarighet. Att välja ut vilka led/frågor som i ett senare skede verkligen ska användas som realistiska frågor/svarsled i ett användningsperspektiv tillhör dock den mer övergripande frågegenereringsuppgiften och behandlas inte direkt i detta projekt.

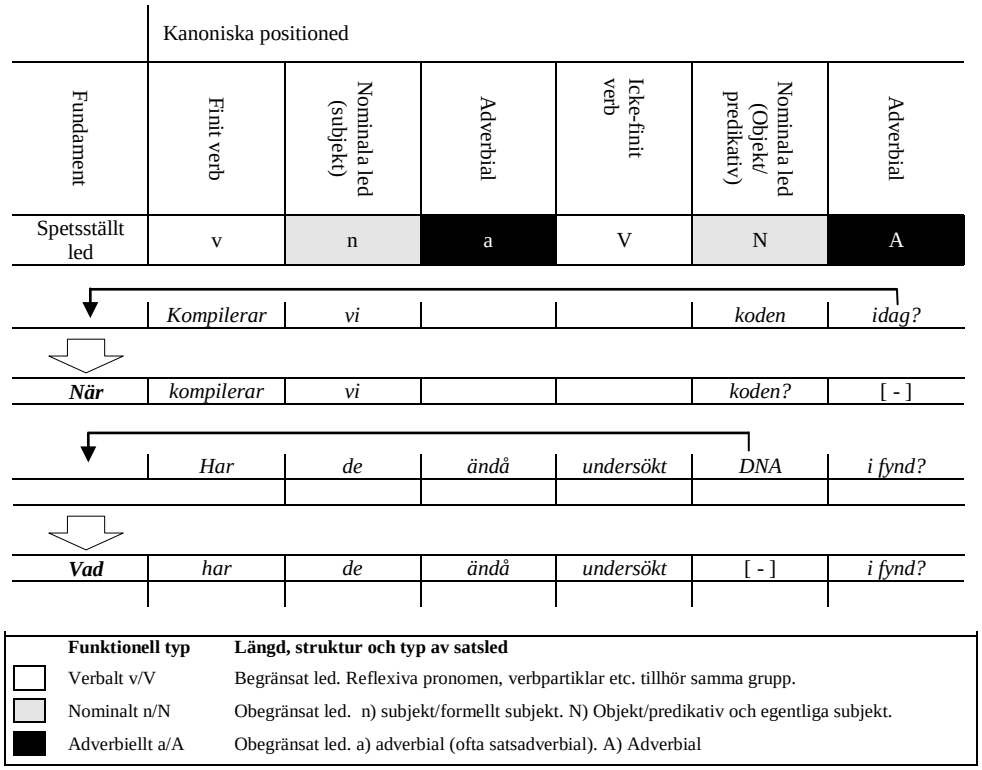
Tack

När det gäller aktuella resurser har jag fått hjälp att finna lämpliga versioner av materialen SALDO och SweFN av Markus Forsberg. Jag har fått svar på vissa funderingar rörande informationsstruktur i detta sammanhang av Maia Andréasson och andra forskare vid institutionen. Maria Toporowska Gronostaj har likaledes svarat på mina frågor om lexikala resurser. Dimitrios Kokkinakis har som alltid varit behjälplig med olika utarbetade lexikala resurser, här för viss namnigenkänning (*named entity recognition, NER)* av olika slag. *Tack också till de många andra medarbetarna vid Språkbanken* för allt som dyker upp under en sådan här period. Jonatan Uppström, Leif-Jöran Olsson och Olof Olsson bistod t.ex. när möjligheten undersöktes att använda Språkbankens webbtjänster dynamiskt till programmet. Detta blev inte fallet i implementationen just här men innebar en intressant möjlighet.

Innehåll

1 Inledning: Allmänt om projektet och rapporten

Denna rapport gäller projektet Adverbialkarakteristik för praktisk informationsextraktion i svensk text. Projektet bedrevs huvudsakligen i januari-mars 2012 på Språkbanken och därefter i vidareutvecklad form under juni samma år.[1] Här beskrivs vad som framför allt har varit ett praktiskt programmeringsprojekt med syftet att automatiskt ge frågeledsmotsvarigheter för fulla satsled i svensk digital text, speciellt adverbialled. Den färdiga funktionaliteten möjliggör att genom ett prototypprogram skriva in adverbial och returnera resultatet: det motsvarande frågeledet; exempelvis *varför,* tillsammans med en viss grafisk och textmässig information som delvis klargör vad som ligger till grund för resultatet.



Tabell 1 Proceduren att spetsställa varje obegränsat led och byta ut det mot frågeled (t.ex. *hv*-ord) följer samma mönster genomgående och tydliggörs med fördel i Diderichsens satsschema (Diderichsen, 1946). De två utgångssatserna här kan t.ex. ha varit *Koden kompilerar vi idag* resp. *De har ändå undersökt DNA i fynd*. Rent tekniskt kan frågor skapas, som här, genom att en V1-frågeform (ja/nej-fråga) först skapas genom att det led som inledningsvis finns i fundamentet placeras på sin kanoniska position, varefter varje närvarande obegränsat led behandlas, dvs. spetsställs och byts ut. I exemplet visas bara hur ett adverbial och ett objekt genomgår denna process men den sker alltså även (i den mån det är möjligt) för de andra obegränsade leden *(vi, koden, de, ändå, i fynd).*

[1] Centre for Language Technology (CLT) finansierade januari-mars. I juni kom finansieringen från Institutionen för svenska språket.

Den föreliggande rapporten tar framför allt upp de metoder som är nödvändiga för att utföra uppgiften och försöker belysa skillnaden mellan en analys för detta praktiska ändamål och den rika teoretiska adverbialbeskrivning som återfinns i litteraturen, och som drar nytta av betydelsesidan hos adverbial, en sida som i en mening är naturligt frånvarande i hela den programmerade mekaniska processen.

I Tabell 1 klargörs den metod som svensk frågegenerering generellt kan göras med, enligt Wilhelmsson (2010), hädanefter: *Avh*. Det har visat sig att just utbytet av de spetsställda leden (här de ganska okomplicerade *idag* till *när*, resp. *DNA* till *vad*) har blivit en stor felkälla. Det aktuella projektet gäller processen vid detta utbyte, speciellt för adverbialled.

Projektet och rapporten har egentligen inte haft för avsikt att uppehålla sig speciellt mycket vid rent teoretiska grammatiska aspekter, förutom där det faktiskt är nödvändigt. Med den aktuella praktiska uppgiften för ögonen visar det sig att den oftast betydelsegrundade kategorisering och beskrivning som återfinns i litteraturen ändå har betydelse för den tekniska uppgiften. Ett avstamp tas här i den traditionella grammatiken, men det blir tydligt hur många av de kategorier som vanligen förekommer, t.ex. *tidsadverbial*, motsvaras av ett flertal olika strukturer, och det är just strukturell analys som är möjlig att genomföra i parsrar generellt och i den parsningsliknande process som ingår i det aktuella programmet – dvs. att känna igen aktuella led som PP, bisatser osv.

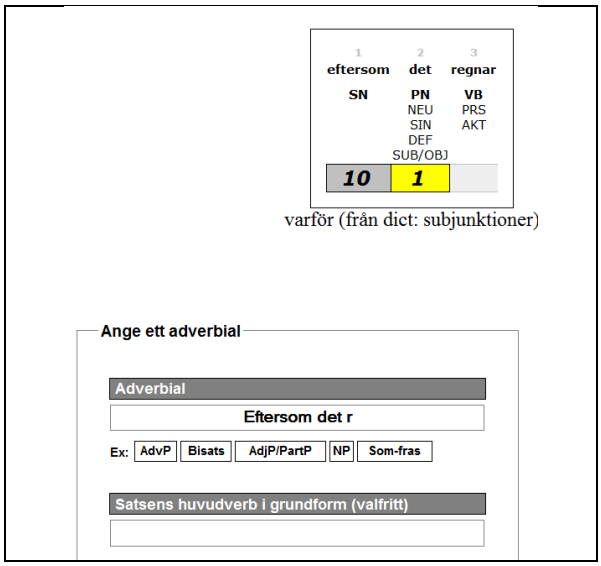
När det gäller använd terminologi och grammatikteori kanske det ska påpekas att användningen är pragmatisk och att arbetet alltså har företrädesvis praktiska syften. Rapporten är ogranskad. Frågor och kommentarer om detta och annat i denna rapport tas gärna emot.

På webben

Delar från projektet kan finnas tillgängliga på webben, i skrivande stund på följande URL:[2]

- www.ling.gu.se/~kw/applications/adverbialkarakteristik/index.htm

[2] I ett senare läge kan prototyper istället göras tillgängliga via författarens aktuella hemsida.



Figur 1 En tidig version av den webb-baserade testprototypen visar resultat och något av vad som ligger till grund för valet.

Denna prototyp har ett avskalat gränssnitt och syftar till att åskådliggöra utvecklad funktionalitet. (Gränssnittet liknar andra gränssnitt som har använts under projektet för att då dynamiskt skriva till nya regler vid åsynen av felaktiga val.)

Rapportens disposition

- Avsnitt 2, Ursprung och syfte med adverbialkarakteristik, bidrar med en beskrivning av uppgiften utifrån sitt ursprung i en tillämpning. En fråga som tas upp är varför uppgiften inte tidigare genomförts. I detta avsnitt finns en koppling till frågedefinitionen i *Svenska Akademiens grammatik*, (1999), hädanefter *SAG*. Kapitlet tar också upp några undantagna ledtyper som ibland har kallats adverbial i litteraturen.

- I Avsnitt 3, Betydelsegrundad adverbialbeskrivning, tas kortfattat upp de betydelsegrundade adverbialkategorier som återfinns i den traditionella grammatikbeskrivningen. Syftet med avsnittet är delvis att kontrastera mot det följande avnittet som tar upp den nödvändigt strukturgrundade ansatsen.

- Avsnitt 4, Strukturgrundad adverbialbeskrivning, tar upp det här avgörande tillvägagångssättet att kategorisera led strukturellt, och därifrån ha speciella metoder för olika strukturtyper som NP-adverbial, PP-adverbial m.fl. Detta är förutsättningarna för att skriva regler för att mappa ledförekomster till frågeled. Detta sker för adverbialen utan att

uttryckligen avgöra vilken betydelsegrundad kategori (t.ex. rumsadverbial) det är fråga om.

- Avsnitt 5, Praktisk adverbialtypsidentifikation, behandlar metoden som används för att i en implementation kunna ordklasstagga indata, avgöra strukturtyp, identifiera huvudord och rektionshuvudord, användning av lexikala resurser och grundformsfunktionalitet.

- Avsnitt 6, Vidare steg med icke-PP-formade adverbial, beskriver hur mappningsförhållandena ser ut för flera strukturtyper som AdvP, PartP, bisatser, NP m.fl.

- Avsnitt 7, Vidare steg med PP-formade adverbial, beskriver det stora arbetet med de många olika grupperna av PP-adverbial. PP-strukturen kräver betydligt fler fall och typer än andra adverbial där huvudordet självt ofta är direkt vägledande för motsvarigheten i frågetyp.

- Avsnitt 8, Om det tekniska utförandet och regelskrivning, beskriver de inblandade programmen, framför allt webb-implementationen i JavaScript och praktiska detaljer. Avsnittet tar bl.a upp hur det befintliga programmet kan förbättras ytterligare i framtiden, för att åstadkomma ännu bättre korrekthet. I princip är det fråga om ett liknande arbete som det som genomförts, dvs. att fel identifieras och att nya regler placerar in undantag mappningsprocessen.

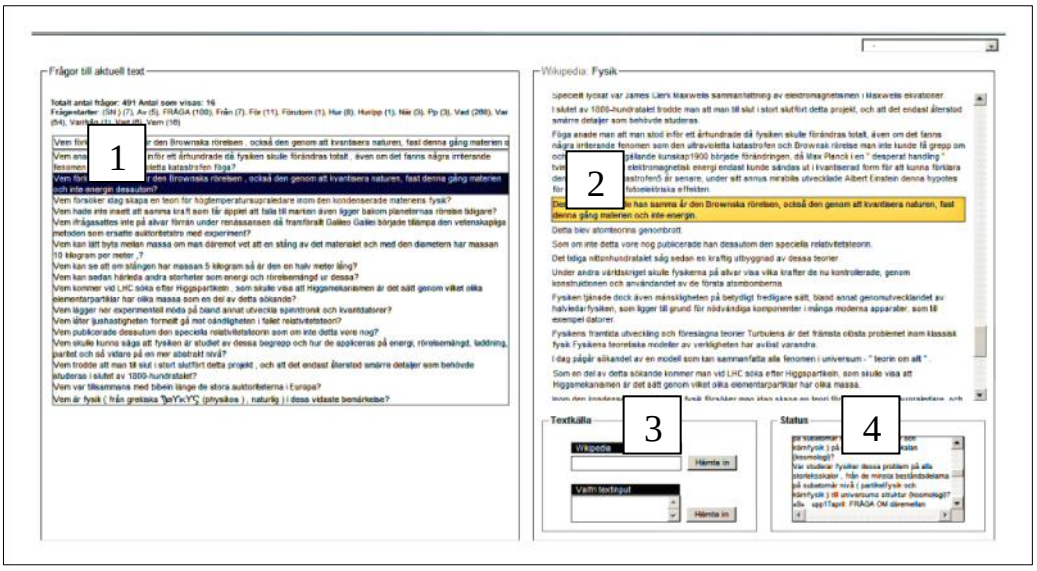
- Rapporten avslutas med Referenser och Appendix.

2 Ursprung och syfte med adverbialkarakteristik

Projektets syfte har alltså varit att från ett informationsperspektiv undersöka svenska adverbial genom att implementera en omfrågningsfunktionalitet så att indata i form av ett användarinskrivet svenskt adverbial *(vid huset)* ges motsvarande frågeled *(var).* Inte alla svenska adverbial kan på detta sätt tydligt utgöra svar på en *hv*-fråga. De semantiskt grundade kategoriseringarna som förekommer i exempelvis SAG, pekar ut några grupperingar med annorlunda omfrågningsegenskaper. Satsadverbial *(inte, ju, såvitt jag vet)* är ett sådant exempel, som enligt källan istället närmast fungerar som svar på ja/nej-frågor (se vidare i avsnittet om betydelsegrundad adverbialbeskrivning).

Uppgiftens ursprung

Det var tillämpad forskning i form av en prototypimplementation av automatisk frågegenerering för svenska som väckte frågorna som detta projekt avser att kasta ljus över. Detta var den sista av tillämpningarna av schemaparsning i Avh. I programmet för automatisk frågegenerering som skapades genom schemaparsning av svensk digital text, Figur 2, blev frågeordsvalet[3] alltså en speciellt felbenägen process, inte minst alltså för adverbialen. De andra obegränsade leden i en funktionell satslösning, de nominala (subjekt, objekt/predikativ), omfrågas med en till synes mindre uppsättning frågeled, ofta *vad* och *vem*. Adverbialen hade en från början tydligt komplex frågeledssida. Idén om frågegenerering för svenska i Avh uppkom oberoende av liknande forskning för engelska, vilken kom till kännedom för författaren senare. För svenska finns, så vitt känt av densamme, inte någon liknande systematisk forskning tidigare.



Figur 2 Användargränssnittet i programmet för frågegenerering (från Avh) upptas huvudsakligen av formulär för frågeval (t.v.) och själva texten (t.h.).

1. Autokompletterande inputfält för val av fråga
2. Texten som hela tiden visas för användaren, där svaret på en vald fråga scrollas fram och markeras
3. Val av artikel i Wikipedias textdatabas eller annan textinput
4. Statusruta för diverse information under körning

[3] Det som här benämns *frågeord* eller *frågeled* är det resultat som föreliggande uppgift ger för svenska adverbial. Det kan ha formen av ett *hv*-ord som *när* men även t.ex. pied piping *(till vem)*.

I implementerade informationssystem mot digital text som försöker ge svar på användarställda frågor (*question answering*) kan mycket skilda metoder användas. Ett system kan t.ex. fungera så att en användare kan ställa en egenproducerad fråga, t.ex. en *varför*-fråga, varefter programmet under körning försöker leta efter alla de möjliga textsegment i texten som kan utgöra ett svar på denna fråga. Exemplet med en *varför*-fråga visar en komplicerande aspekt: Somliga frågetyper kan ha och har ofta mycket komplicerade möjliga textsegment som svar. En *varför*-fråga som förekommer i FAQ-sammanhang har troligen ett svar som är formulerat i en mängd huvudsatser. Men en svarsform kan även vara ett enda adverbialled, som den för *varför*-frågetypen dedikerade *eftersom*-bisatsen.

Medan det alltså är en mycket komplex uppgift att mappa frågor mot alla möjliga svarsformer och identifiera dessa i en text så är avgörandet av motsvarande frågeled för varje satsled, och uttrycklig generering av besvarade frågor, enklare. Det är denna möjlighet som ligger till grund för den aktuella ansatsen. Det visar sig att om utgångspunkten är satsled (svar på frågor) så är dessa ofta ’en-värdiga’ i fråga om möjliga frågemotsvarigheter, en PP som innebär *i + plats* kan t.ex. konsekvent mappas enbart till *var*-frågor. – Utgående från satsleden på detta sätt finns det en chans att arbeta mer systematiskt för att identifiera informationsinnehållet i frågetermer. Detta är syftet med den aktuella forskningen.

Det går att tänka sig hur ett välutvecklat system som kartlägger en texts ingående informationsinnehåll på syntaktiska grunder kan möjliggöra för en användare att avgöra om en viss fråga över huvud taget kan besvaras av en texts informationsinnehåll, var i texten ett eventuellt svar på någon tänkt fråga måste finnas, om det alls är närvarande.

Det aktuella projektet utgör en deluppgift inom frågegenerering och rör enbart utbytena av adverbialled till frågeled, dvs. inte generering av fullständiga frågor. Adverbialled måste i prototypen som hör till detta projekt skrivas in av användaren. I prototypen finns även möjlighet att uppge ett huvudverb i grundform. Om adverbialet är en PP och verbet visar sig valensmatcha den inledande prepositionen kan programmet ange en frågeform som är anpassad för en ’prepositionsobjektstolkning’.

En helt ny uppgift?

Uppgiften som alltså uppkom i en tillämpning kan förmodligen ses som tillhörande den språkvetenskapliga grundforskningen (vilken ändå kan ha en

viktig roll i regelbaserade språktekniska tillämpningar). Det kan ses som en grundläggande uppgift att avgöra vilken kategori som ett visst adverbial tillhör och vilka de motsvarande frågeleden är. Utan frågegenerering i sikte blir emellertid denna uppgift kanske inte särskilt lockande att ta sig an – eller hur kommer det sig att ingen större forskning har försökt utreda denna relativt iögonfallande frågeställning? Medan det egentliga behovet uppstår just i språktekniska tillämpningar är uppgiften väl intressant även i ett allmänt språkvetenskapligt grundforskningsperspektiv?

Svaret på frågan om varför uppgiften inte ägnats speciellt mycket uppmärksamhet förses med flera svarsförslag i denna rapport. Som redan nämnts blir uppgiften ofrånkomlig i en implementation av frågegenerering. Frågegenereringen bygger på en syntaxanalys med ett visst format: nämligen med grammatiska funktioner där leden identifieras på rätt satsnivå och med sina hela sträckningar (framför- och efterställda attribut), en egenskap som faktiskt inte förekommer med bra korrekthet i speciellt många parsrar som hanterar svensk text. Därmed är frågan inte aktuell speciellt ofta av praktiska skäl.

Men som svar på frågan om uppgiftens eventuella förbigångenhet finns också dess varierande och sammantaget ganska omfattande krav i form av digitala lexikala resurser och i form av algoritmer för t.ex. huvudordsanalys. Generellt krävs för adverbial metoder för att avgöra huvudord och rektionshuvudord (för PP). Denna rapport visar en användbar metod för dessa uppgifter.

Ett tredje svar på frågan hänger samman med att några typer av adverbial för att omfrågas (eller för att deras komplement ska kunna omfrågas), innebär en nödvändig närkontakt med den stora ordklassen substantiv och deras semantiska aspekter. Substantiv kommer framför allt in i bilden i hanteringen av PP-adverbial och de mer sällsynta NP-adverbialen. Från ett praktiskt perspektiv krävs för rationellt arbete också funktionalitet för avgörande av ordgrundformer.

Somliga frågor blir troligen inte upptagna i forskningen då de inte verkar innebära en tillräcklig intellektuell utmaning. Detta är i förstone något som verkar gälla för den aktuella uppgiften: det är en uppgift som verkar ha en rent uppräkningsmässig natur. Denna typ av uppgifter som rör en mappning, om än i ett vidsträckt, mångfacetterad område, har i språktekniskt perspektiv allt oftare blivit föremål för statistiska och maskininlärningsbaserade metoder. För att genomföra så kallad ’supervised machine learning’, dvs. en metod där en uppmärkt träningsdata begagnas för att framställa ett program som kan lära sig att dra slutsatser även i nya okända fall, behövs alltså denna uppmärkta träningsdata – och sådan data som kopplar led till frågor existerar nu inte. Återstår så (halv-) manuellt arbete för att åstadkomma en sådan mappning direkt, eller att vidareutveckla automatiska metoder utifrån. Den mängd

manuellt arbete som verkar vara nödvändig för att göra en välfungerande mappning visar sig i detta projekt vara stor.

Bland de ovan nämnda skälen för eventuell förbigångenhet återfinns inte några tankar om att uppgiften skulle innehålla några omöjliga eller teoretiskt oklara fall i mappningen. Men arbetet har visat att sådana fall finns, i det ”totala” perspektiv som innebär att nästan alla adverbialled beaktas, visas att uppgiften på sina håll är behäftad med rent teoretiska svårigheter. Denna typ av oklarheter, ifall led alls kan omfrågas (t.ex. *slutligen*), eller vilket av flera sätt som bör väljas *(till 73 kandidater: till vad, till vem, till hur många kandidater*) har på sina håll gjort att det ibland har varit svårt att säga vad rätt svar (dvs. fråga!) för ett visst adverbial egentligen bör vara.

Att det inte finns någon tydlig föregångare på fältet kan tolkas som att den aktuella frågeställningen inte är speciellt ’het’ i forskningssammanhang. Men som antytts är frånvaron av forskning inte nödvändigtvis beroende på ointresse. Med en helt fungerande lösning öppnas troligen dörrar till ny spännande forskning inom området.

SAGs logiska definition av sökande frågor

SAG har en grundläggande indelning av frågeformer i V1-frågor (prototypiskt *ja/nej*-frågor, vilka kallas rogativa) och frågor som har fundament (’interrogativt led’), *hv*-frågor, vilka kallas kvesitiva. De frågor som inte är ja/nej-frågor kallas gemensamt för *sökande frågor*.

I *SAG* beskrivs sökande frågor, om t.ex. adverbialled i uppgiften här, i ett logiskt perspektiv på ungefär följande vis: en *hv*-fråga innebär en fråga om vilka premisser som krävs för att en viss proposition ska få positivt sanningsvärde. Med andra ord kan en sådan fråga om ett adverbial beskrivas som ’*hv*-led (adverbial) + resten av propositionen’ – t.ex. *<u>När</u> kommer Lisa hem?* Frågeställaren förutsätter att resten av propositionen *(Lisa kommer hem)* under vissa förutsättningar renderar ett positivt sanningsvärde och använder *hv*-frågan för att söka efter just efter dessa förutsättningar. I den aktuella uppgiften är det alltså sökande frågor och primärt de som svarar mot fulla satsled som behandlas.[4]

[4] SAG beskriver hur de båda formaspekterna, V1 eller *hv*-fråga, inte entydigt kan svara mot skillnaden i funktion. Dvs.: efterfrågas sanningsvärdet (typiskt V1) eller efterfrågas de premisser som krävs hos ett visst ledslag för att ge en sann utsaga? Ibland används de i praktiken på mer komplicerade sätt.

En kritik

Uppgiften frågeledsgenerering här, som en del i den överordnade processen frågegenerering, kommer alltså från ett sammanhang då samtliga obegränsade funktionella huvudsatsled (subjekt, egentligt subjekt, objekt/predikativ och adverbial) ses i ett informationsperspektiv. Det betyder att den fullständiga mängden frågor per huvudsats kunde vara som i nedanstående exempel, en ”total” ansats.

Han spelade den andra matchen på lördagskvällen

Vem spelade andra matchen på lördagskvällen?
Vad spelade han på lördagskvällen?
När spelade han andra matchen?

Detta är verkligen det resultat som en byggd tillämpning av frågegenering för svenska enligt beskrivningen ovan siktar in sig på att ge.[5] Uppgiften har ibland beskrivits som syftande till att ge samtliga frågor en text kan sägas besvara. Vad detta ens betyder är oklart. Det finns här en troligtvis befogad kritik mot detta totala perspektiv som kan innebära uppemot samma antal frågor som antalet obegränsade huvudsatsled. I ett prototypsystem (Wilhelmsson, 2011) skapas frågorna på detta icke-diskriminerande sätt. Det blir lätt hundratals frågor för relativt korta texter. Där visades hur i genomsnitt 4.0 frågor per textmening skapades. Användargränssnittet för praktisk användning är byggt med en auto-kompletterande dropdown-meny som är tänkt att leda till att en befintlig besvarad fråga åt gången väljs (matchande tillgängliga frågor visas när användaren börjar skriva). Problemet är, förutom det faktum att den aktuella frågeformuleringen och ordval kan göra en viss fråga svår att hitta, det stora antalet frågor. Resultat från prototypen för svenska, liksom liknande försök för engelska, visar att de grammatiskt riktiga frågorna inte heller nödvändigtvis är sådana som en användare vill ställa mot en text. En stor andel, av de frågor som är grammatiskt korrekta och i princip kan sägas bli besvarade, är helt enkelt inte relevanta eller användbara nog.

För engelska språket där frågegenerering har växt till sig mer som forskningsfält förekommer också något som liknar den beskrivna totala ansatsen och är benämnd *overgenerate and rank*. Processen som beskrivits kallas alltså uttryckligen för *övergenerering* och när olika former av olämpliga frågor

[5] När det gäller samordnade finita verbfraser på huvudsatsnivå så transformeras dessa till huvudsatsform med subjekt i programmet genom att ärva huvudsatssubjekt därifrån det senast förekommer: *Kalle spelade i morse och gav Lisa en present → Kalle spelade i morse, Kalle gav Lisa en present:* fem möjliga ledfrågor.

(ogrammatiska, irrelevanta etc.) plockats bort och rankats kan så lite som en tiondel av frågorna återstå.

I detta läge kan språkvetenskapliga teorier om informationsstruktur ha något att tillföra för att möjligtvis kunna säga något om vad som relevant i en helt okänd text. En hypotes utan egentliga följder i det aktuella arbetet är att *remaled* eller *fokusled* som kommer med ny information är mer benägna att utgöra relevanta frågesvar än något typiskt gammalt, som t.ex. ofta kan kännas igen i text genom formen av ett fundamenterat pronomen. Hittills saknas implementationer, åtminstone för svenska, av en teoretiskt adekvat fungerande uppmärkning med informationsstatus. Dessvärre finns än flera olika modeller i omlopp när det gäller begrepp inom informationsstruktur, även om varje modell med sina anhängare har välfungerande definitioner. Det vore ett intressant experiment att försöka utesluta åtminstone somliga frågor som på ett tydligt sätt härrör från uppenbara bakgrundsled i processen, för att se om det därmed utesluter ’mindre användbara frågor’.

De flesta led som förekommer *kan* troligen vara rematiska, omfrågningsmässigt relevanta led under rätt förutsättningar och uppgiften mappning till frågeled här tar ju inte hänsyn till sådant som möjligen indikerar aktuell informationsstatus för ett visst led. Därför berör kanske inte kritiken om producerade frågeleds relevans just mappningsprocessen av de enskilda leden utan handlar om lämplighet/olämplighet av generering av hela frågan på grundval av fler aspekter som ledets aktuella placering och de andra förutsättningar som avgör dess informationsstatus. Den faktiska genereringen av fulla frågor, enligt Tabell 1 och Figur 2, och fastställandet av vilka frågor som är relevanta ligger bortom denna rapports domäner.

En tidig indelning av adverbialslag

I ansökan till projektet visades nedanstående tabell, vilken är en preliminär indelning från ett slags operationaliserbarhetsperspektiv. Det är fråga om grupper av adverbial sådana att ett adverbial under analys antas korrekt kunna klassificeras till rätt grupp. De olika kategorierna är företrädesvis strukturgrundade, men det finns här kategorier som *satsadverbial*, vilka har många möjligheter att anta olika former och utgör en grupp med många former och som behandlas speciellt beroende på att de har speciella frågemotsvarigheter, nämligen valet *’ingen motsvarighet’* för de flesta satsadverbial.

Grupp	Beteckning	Exempel	Huvudsakliga strukturtyper	Förhållande till frågeord
A	Satsadverbial och liknande utan enkel frågeordsmappning	*Ej, dock, i så fall*	AdvP, PP	-
B	Adverb och participfraser med möjligt frågeordsförhållande	*Lika snabbt*	AdvP, PartP	HO eller delar av frasen
C	Bisatsadverbial, inkl. *som*-satser	*Eftersom… Som de trodde…*	Bisatser	Bisatsinledare – *hv*-ord
D	NP-formade adverbial	*Denna gång*	NP	Huvudord – hv-ord. Dessa led kan alltid fungera 'nominalt' i vissa strukturer
E	Attribut på satsnivå	*, vilket gladde oss.*	Vilket-sats	-
F	*Som*-fraser	*Som Kalle, Som på 1990-talet*	Som-satser	*Som* + rektionens *hv*-ord
G	'Prepositionsobjekt'	*[lyssnar] på musik*	PP	(oftast inte *hv*-ord), enbart pied piping eller rektionsframflyttning
H	'Normala' PP-formade adverbial	*I skogen, med Paris*	PP	Mångfacetterat, se nedan [intern referens]

Tabell 2 En tidig indelning av svenska adverbialtyper grundad på struktur, omfrågningsbarhet och till viss del betydelse. Från ansökan till det aktuella projektet.

Som ovanstående tabell visar har svenska betydelsegrundade adverbialtyper mångskiftande syntaktiska strukturer. Redan från början kan konstateras att somliga adverbial, i föreliggande uppgift, t.ex. vissa adverbfraser *(här: var*, *nu: när)* har en okomplicerad 'ren mappning' från huvudord till lämpligt frågeled. Adverbial i form av prepositionsfraser låter sig däremot inte mappas till frågeledsmotsvarighet utan användning av en större samling språktekniska resurser. Det gäller identifikation av huvudord och rektionshuvudord (genom ordklasstaggning och s.k. *ranger*), grundformsfunktionalitet för rektionshuvudorden samt med något medel klargörande om rektionshuvudords natur (t.ex i pied piping-frågor för prepositionsobjekt) – för att idealiskt kunna skilja t.ex. *för vad* från *för vem*.

I denna rapport används termen *pied piping* så att när ett adverbial får pied piping som lösning så är även en systerkonstruktion: *rektionsframflyttning med strandad preposition* giltig. Det innebär att varje gång en frågeform presenteras som pied piping är den utbytbar mot denna form, *På vad…* är alltså utbytbart mot *Vad… på*. I praktiken är troligen denna senare frågetyp den vanligare. I rapporten används dock pied piping *'På vad'* som något av ett paraplybegrepp. När ett adverbialled tilldelas en pied piping-lösning så betyder det att den reella frågan kanske snarare bör bli den sönderdelade versionen – men alltså inte den andra kategorin; ensamma *hv*-led motsvarande hela ledet (som *var*). Rent programmeringstekniskt är det enkelt att producera den sönderdelade varianten

(Vad sitter de på?) från pied piping-varianten *(På vad sitter de?)*. Det är dock praktiskt att spara dessa frågeled i pied piping-form, eftersom det är en kompakt och sammanhängande form. Hädanefter i detta arbete kommer med pied piping menas *pied piping, eller den ekvivalenta systerkonstruktionen: rektionsframflyttning med strandad preposition*.

Undantagna ledslag som ibland har kallats adverbial

Somliga ledslag behandlas i vissa (teoretiska) arbeten som adverbial men har inte beaktats i det aktuella frågeperspektivet. Det beror i vissa fall på att de saknar frågemotsvarigheter men även på att de över huvud taget inte känns igen och räknas in bland adverbialen i den strukturanalys som görs här, vilken i sin tur grundar sig på ordklasstaggningen i Stockholm Umeå Corpus 2.0. (Ejerhed, Källgren, & Brodda, 2006), hädanefter *SUC*. Märkningen i SUC är, genom korpusens dominanta ställning som träningsdata för ordklasstaggare, en betydelsefull faktor för hur de parsrar som hanterar svenska faktiskt analyserar. Det leder t.ex. till att participfraser urskiljs som en särskild frastyp och att adjektivfraser kan tolkas som adverbial.

V1-formade villkorsadverbial

Frågeformade villkorsadverbial (*Regnar det går vi in*) har som andra villkorsadverbial (t.ex. *om*-bisatser) ett möjligt frågeled i det ganska formella *”under vilka förutsättningar”* (det är enkelt att åstadkomma) men dessa hanteras för närvarade inte.

Partikeladverbial

’Partikeladverbial’ jämförs i t.ex. *SAG* med andra adverbial. Som namnet antyder finns en släktskap som dock definieras bort här i den aktuella ansatsen. Definitionen på partikeladverbial är att de inleds med partikel. Dessa är, i flera senare grammatiker, alltid betonade sådana och skiljer sig en del från vanliga adverbial (t.ex. PP-adverbial); detta arbete som utgår från ordklasstaggningen i SUC 2.0. känner i text igen partiklar genom ordklasstaggen partikel (PL) och i schemaparsningen är detta inte ett adverbialled liknande PP utan denna partikel är ett av de begränsade (bounded) leden som t.ex. verb. Komplementet betraktas som ett eget (ofta nominalt) led: *slå i en spik*.

Varslande adverbial

I lundaprojekten om tal- och skrivsyntax (Teleman, 1974), men i övrigt sällan,[6] förekommer benämningen *varslande adverbial* (VA) för första ordet i ordpar som *varken – eller* eller *både – och*. I SUC 2.0. är dessa (parvisa) konjunktioner (KN) märkta precis som de andra orden i ordparen, dvs. som konjunktioner, och inte som adverbial.

Attribut på satsnivå

I det fältgrammatiska perspektivet faller led av typen *", vilket roade oss"*, och som hänför sig till hela satsen, utanför de mest centrala sats- och ledformerna. I schemaparsern märks leden upp som adverbial, närmast genom egenskapen att vara optionella satskonstituenter. Både placerings-, form- och betydelsemässiga restriktioner gör dem otypiska. I en språkteknisk applikation som eftersöker propositioner vore det förmodligen rimligast att transformera dessa till egna huvudsatser *(detta roade oss)*, vilka kunde generera egna frågor utifrån sina ingående led.

Gränsfall: fokuserande adverbial

Satsadverbial som är bestämning till enskilt led snarare än till hela satsen benämns ibland fokuserande adverbial *(Jag nästan sprang)*. Om dessa analyseras som normala adverbial; dvs. om *jag nästan sprang* betraktas som ekvivalent med *jag sprang nästan* så kunde ledet i princip hanteras i frågegenereringsprocessen som de andra. (Emellertid är dessa led alltså ofta satsadverbial och har därmed oftast inte någon enkel frågemotsvarighet.)

Dubbeladverbial

"Dubbeladverbial är en kombination av två (eller flera) adverbial som tillsammans utgör ett primärt satsled [...]" (SAG, *Fraser*, s. 441).

Dubbeladverbial består av flera adverbial som båda kan vara omfrågningsbara: *från Gambia till Tanzania.* Om ett dubbeladverbial bestående av två PP, som i exemplet, anges som ett adverbial i programmet som här skapas så kommer det att behandlas som ett enda adverbial. I praktiken kommer alltså den andra

[6] En Internet-sökning ger väldigt få träffar för denna term som dock kan förekomma i annan litteratur.

prepositionsfrasen att bortses ifrån. Det beror på en här ouppklarad svårighet att skilja dessa från PP-adverbial som består av två PP där det andra är attribut till det föregående: *till en man från Bern*.

(En del av) de bundna adverbialen i SAG

En viktig skillnad mellan *SAG* och den analys som sker i detta projekt, även för schemaparsern är kategorin bundna adverbial. I *SAG* (s. 440) räknas NP och nominala bisatser ibland som bundna adverbial:

> Taket kostar mig 40000 *att reparera*
> Våra chips går inte *att äta*

Denna undergruppering behandlas dock inte som adverbial här. De analyseras i schemaparsning som attribut eller nominala led och verkar inte ha någon enkel omfrågningsaspekt.

3 Betydelsegrundad adverbialbeskrivning

Adverbialindelningen i litteraturen är som nämnts i allmänhet semantikorienterad. Satsadverbial bildar en speciell kategori, medan de övriga kan kallas innehållsadverbial och delas in på ganska varierande vis. I SAG behandlas innehållsadverbialen i kapitlet om verbfrasen. Det följande avsnittet är en mycket kortfattad återblick på denna traditionella kategoriindelning. Från början ska konstateras att denna indelning alltså inte är praktisk att arbeta efter: vad som kan identifieras inledningsvis av en parsningsprocess som i det aktuella programmet är vilken struktur ett led har (dvs. PP, bisats el dyl.) – inte vilken betydelsegrundad kategori (t.ex. rums- eller satsadverbial) den tillhör. Avsnittet härefter beskriver det ovanligare synsätt som är avgörande i implementationssammanhanget: strukturgrundad adverbialbeskrivning.

Som visas i tabellen nedan slår den betydelsegrundade kategoriseringen tvärs över den strukturgrundade. Att automatiskt avgöra vilken betydelsegrundad adverbialkategori ett visst adverbial har automatiskt är härmed svårt direkt från indatan varifrån struktur är vad som avläses. Den aktuella uppgiften innebär dock en ännu mer finkornig beskrivning, frågeled ska avgöras, och det är möjligt att skapa en procedur som givet dessa svar, t.ex. *när/hur länge* också då anger att det i det fallet är fråga om kategorin tidsadverbial.

	AdvP	PP	Bisats	NP
Satsadverbial:	*inte*	*för all del*	*om jag minns rätt*	*min själ*
Rumsadverbial:	*där, ditåt*	*till Kina*	*innan Kiruna*	*20 meter*
Tidsadverbial:	*nu*	*i denna stund*	*när vi sågs*	*ett tag*
Andra:	*så där*	*på detta sätt*	*eftersom det går*	

Tabell 3 De adverbialkategorier som förekommer i litteraturen och som är betydelsegrundade har många olika motsvarande strukturformer.

Satsadverbial

Satsadverbial har en särställning som adverbialtyp och behandlas som egen gruppering ofta mer utförligt än alla andra adverbial. I *SAG* utgör kapitlet om satsadverbial hela 120 sidor. Till skillnad från de övriga adverbialen, (innehållsadverbialen) som tas upp i verbfrasens beskrivning, finns satsadverbialen i en egen kategori genom sina speciella betydelse- och placeringsaspekter. Så mycket skiljer sig dock dessa adverbial inte från andra grupper strukturellt och placeringsmässigt – om de ska kännas igen i programmet behöver det ske genom uttryckliga matchningsregler.

Att klargöra de frågeord som motsvarar eller möjligen motsvarar adverbialtyper är en uppgift med relativt klara lösningar. Speciellt innebär den relativt tydliga förhållanden jämfört med uppgiften att korrekt kategorisera exempel i semantiska grupperingar. Adverbial och satsadverbial sorteras semantiskt på ett flertal sätt i litteraturen. En helt okomplicerad taxonomi verkar saknas.

Prototypiskt har satsadverbial betydelsen av en satskommentar som hänför sig till hela satsen, men ibland till enskilda led (t.ex. *bara*). Satsadverbial får en speciell betydelse i parsning med applikationer där de strukturellt inte enkelt kan skiljas från andra (adverbial-)led vid identifikationen utan lyder speciella regler för gruppering och framflyttning. Frågan i detta sammanhang är huruvida satsadverbialen är omfrågningsbara, ett negativt svar skulle med fördel leda till att grupperingen utesluts från den föreliggande processen.

Det första som kan konstateras är att en grupp satsadverbial, de modala, visst fungerar väl som frågesvar, i fallet ja/nej-frågor: *Har de åkt? Kanske. ([De har] kanske [åkt]).* I detta sammanhang är det emellertid de sk. *sökande* frågetyperna (*hv*-frågor/frågeordsfrågor) som främst beaktas. Här blir den betydelsegrundade kategoriseringens svagheter tydliga. *Aldrig* är t.ex ett satsadverbial i grupperingen nekande satsadverbial, men det kan vara ett rimligt frågesvar på *när* likt somliga tidsadverbial.

Det i SAG framförda satsadverbialtestet är elegant klargörande, men tyvärr inte alltid fullständigt enkelt att använda, eller heltäckande. Det består av två komponenter och begagnar sig av testarens språkliga intuition:

1. Karakteristiskt för de flesta satsadverbial är att de kan parafraseras så att den sats (S) som innehåller ett satsadverbial (SA) kan göras till underordnad sats i en satsfogning där satsadverbialet står i den överordnade satsen enligt mönstret:
   *”Det är SA så att S”. (Det är [inte/*idag] så att Erika åkt.)*

2. Samtidigt ska satsadverbial *inte* kunna brytas ut ensamma ur den sats de ingår i:
   ** Det är {troligen/således/åtminstone…} som Erika åkt.*

   (Fritt efter *SAG*, Band 4, s 85)

   Testet kommer att korrekt identifiera de flesta satsadverbial mycket väl:

- Det är inte så att Erika åkt. / * Det är inte som Erika åkt.
  Det är efter vad det verkar så att Erika åkt. /* Det är efter vad det verkar som Erika åkt.

När det gäller typiskt satskommenterande adverbial *med all sannolikhet, faktiskt, som sagt* m.fl. vore den närmsta frågan (förutom V1-frågor) möjligen *hur*, men en *hur*-fråga som besvaras med satsadverbial förefaller också orimligt (*Hur är det Erika åkt? – Efter vad det verkar*).

Med andra ord finns det både eventuellt omfrågningsbara (*aldrig*) och icke omfrågningsbara (*inte*) satsadverbial, när sökande frågor används för att efterfråga information i satsen. Det ska nämnas att i längre strukturer t.ex. PP finns det även möjlighet till vad som kan kallas *attributfrågor (Enligt min mening – Enligt vems mening?).* Satsadverbial är på det hela taget en gruppering

som ofta innebär *'inget'* som resultat. Denna lösning är i sig också ett avgörande och bedöms som korrekt eller fel.

Precis som de andra adverbialslagen är satsadverbial alltså fördelade över många olika strukturfall. I det aktuella programmet identifieras satsadverbial uttryckligen i viss mån eftersom de som sådana oftast har 'ingen frågeledsmotsvarighet' som resultat. (Somliga har dock mappats direkt till denna 'ingen frågemotsvarighet' utan att de uttryckligen märkts som satsadverbial, eftersom det alltså trots allt är frågemotsvarigheten som det relevanta i slutprodukten.)

Innehållsadverbial

När det gäller de andra adverbialslagen kan de t.ex. indelas efter typ på följande vis, som i Svenska akademiens språklära, *SAS* (Hultman, 2003), s 239. Dessa adverbial behandlas i allmänhet i processen.

- Tid *(på onsdag)*
- Sätt *(snabbt)*
- Rum *(på tåget)*
- Medel *(med tåg)*
- Orsak eller följd *(av kyla)*
- Följeslagare *(med Kalle/utan Kalle)*
- Mått *(en hel vecka)*
- Agentadverbial *(av Kalle)*

Det är ofta i litteraturen fråga om aningen oklara fall och kategoriseringar som ser annorlunda ut i olika grammatikor. I det aktuella perspektivet bortses från dessa olika (när-) taxonomier då syftet är att gå direkt till frågeled.

Frågeledet hur som enda möjlighet

Frågeledet *hur* förtjänar att tas upp redan här. Frågeledet *hur* dyker upp som det enda rimliga frågeledet för många fulla adverbialled i denna ansats som i princip behandlar de flesta adverbial. T.ex. de som gäller medel; *med tåg* eller adjektiv/adverbfraser som *glatt, roligt, snabbt* och participfraser får lösningen *hur*. *Hur* anses dock vara inherent vagt och fungerar inte alltid realistiskt; *Hur upptäckte ni lösningen? – Snabbt.*

4 Strukturgrundad adverbialbeskrivning

Från ett automatiserat analysperspektiv sker adverbialbeskrivningen alltså från utgående från struktur. I programmet har ungefär sju olika huvudgrupper fungerat praktiskt som överordnad grovkategorisering. Denna indelning skiljer sig från den skiss som återfinns i projektansökan som mer fokuserar på frågeformen direkt. Den nedanstående används av praktiska skäl i programmet. Undantag som satsadverbial hanteras inte genomgående för sig utan tillhör en allmän samling undantag.

1 AdvP	*oerhört bra, nästan bättre än innan*
2 PartP	*slående, ganska välstekt*[7]
3 AdjP	*Obegripligt*
4 NP	*Ett tag, de första milen, Kl 14, 1943*
5 Bisats	*eftersom det regnade*
6 Som-fras	*som målvakt, som på medeltiden*
7 PP-former[8]	*av ilska, på bilen, på så sätt, gentemot oss, i sak*

Korrektheten i skrivande stund varierar kraftigt mellan de olika av dessa strukturled. Bisatsadverbial är exempelvis en lätthanterlig kategori medan hanteringen av prepositionsfraser har tagit en mycket stor del av projekttiden i anspråk. Klassen PP-adverbial består av ett relativt stort antal ovanliga prepositioner av de uppemot 200 olika prepositionerna som kan finnas i svenska. Det är dock de frekventa slagen *på, i, till* m.fl. som utgör den stora utmaningen med sina många olika delfall.

5 Praktisk adverbialtypsidentifikation

Den föreliggande uppgiften, att tilldela alla förekomster av adverbial en eller – vid rätt tillfällen – ingen frågeledsmotsvarighet, sönderfaller i en räcka delsteg som är olika lång beroende på aktuellt adverbial. De följande avsnitten redogör

[7] När det gäller vilka ord som i praktiken taggas som particip och inte t.ex. adjektiv finns en viss inkonsekvens i SUC, och det är denna taggning av huvudord som avgör strukturtolkning.

[8] Denna grovsortering är i somliga fall, som för particip-fraserna, tillräcklig för att direkt ange frågeord (för participfraser generellt *hur*). En sortering till PP behöver däremot efterföljas av avgörande om den aktuella enheten är satsadverbial, 'prepositionsobjekt', har vad som här kommer att kallas starkt huvudord (som *sätt/vis*, vilket ofta avgör frågeformen), tillhör andra fasta konstruktioner (som *ge vid handen*), är agentadverbial eller slutligen tillhör grupperingen 'vanliga' PP-adverbial.

för hur den nödvändiga informationen om strukturtyp, huvudord och rektionshuvudord hos adverbial kan identifieras.

Ordklasstaggning

Ordklasstaggaren som här begagnas har tidigare använts för schemaparsningen med tillämpningar. Det är fråga om en trigrambaserad Hidden Markov-modell med Viterbi-algoritmen samt vissa egna diskrimineringsregler. Den använder använder s.k. additiv smoothning.[9] Den är tränad på större delen av SUC och märker upp med detta taggset. Ordklasstaggaren som är skriven helt i JavaScript är egentligen det enda exemplet på maskininlärningsbaserad metod här. Metoden räknas även som s.k. dynamisk programmering.

Redan	då	var	trafiken	tät	mellan	stadsdelarna	och	en	bro	med	kapacitet	behövdes	.
AB	**AB**	**VB** PRT AKT	**NN** UTR SIN DEF NOM	**JJ** POS UTR SIN IND NOM	**PP**	**NN** UTR PLU DEF NOM	**KN**	**DT** UTR SIN IND	**NN** UTR SIN IND NOM	**PP**	**NN** UTR SIN IND NOM	**VB** PRT SFO	**MAD**

Figur 3 Enheten hb06a-011 från SUC 2.0 visualiserad med ordklasstaggning.[10]

Ordklasstaggaren är speciell såtillvida att den är helt klientbaserad och har körts lokalt i samband med schemaparsern och applikationer. Den har utvärderats och beskrivs i detalj i Avh.

Identifikation av strukturtyp, huvudord och rektionshuvudord

Gemensamt för samtliga adverbial är att de analyseras syntaktiskt i programmet. De kan sägas parsas frasstrukturellt. Detta sker med en metod hämtad från schemaparsning; Avh. Denna syntaktiska analys klargör strukturtypen.

Identifikationen av huvudordet är tätt förknippad med avgörande av strukturtyp, det är huvudordet som avgör frastyp eller dylikt. Huvudordet är i allmänhet antingen *det inledande ordet* eller vad som i algoritmen nedan kan kallas *det första lokala minimumet* (se beskrivning nedan). För prepositionsfraser anammas hållningen att prepositionen är huvudord men det lokala minimumet är

[9] Denna ordklasstaggare är ett praktiskt delresultat från den egenkonstruerade kursen Computational methods in tagging and chunking, vilken handleddes av Viggo Kann.

[10] De koder som står för information om ordklass och viss annan grammatisk information under texten finns förklarade i Appendix.

där relevant eftersom det generellt blir rektionshuvudordet, vilket för många PP-slag har en lika avgörande roll som prepositionen vid frågeledsvalet.

En algoritm för huvudordbestämning i obegränsade led i svenska

Huvudordsbestämning förekommer ofta parsning men har något varierande roller. Två exempel:

- I shallow parsing med chunkar (Abney, 1991), identifieras ofta segment utan efterställda attribut. Det innebär att det sista ordet per definition blir huvudordet – undantaget PP och bisatser.
- I dependensgrammatisk analys är det frasernas huvudord som först förses med länkar till satsens huvudord osv.

I schemaparsing har huvudordsanalys inte haft riktigt samma fokus på huvudord som på identifikation av det fullständiga ledet på en viss satsnivå inklusive alla attribut. Huvudordsidentifikation har framför allt varit relevant i själva parsningen för att avgöra prominensnivå, dvs. animathet osv. med utgångspunkt i skalor som har undersökts av bl.a. Øvrelid (2008), detta har använts för avgörande av subjekt/objekt. Huvudordsidentifikation har emellertid blivit speciellt viktigt här i avgörandet av motsvarande frågeled.

I likhet med rena ytstrukturparsrar (shallow parsers) grundade på reguljära eller kontextfria grammatiker är den delmetod som förekommer i schemaparsningen (s.k. rangbaserad chunkning) sådan att den identifierar frastyper fram till huvudordet.

Texten, eller det enskilda segmentet, behöver för den metod som här beskrivs antingen vara taggat med ett taggset som det i SUC eller kunna beredas taggning av detta slag. Noga räknat använder den följande rangtilldelningen inte den stora mängden undersärdragsvärden i SUCs taggning, vilken tillsammans med de tjugotal ordklasstaggarna ger ca 150 i praktiken förekommande taggkombinationer.[11]

Indata till själva algoritmen är alltså ordklasstaggad text, företrädesvis med de taggar som förekommer i SUC 2.0 eller med en liknande uppsättning där de nedanstående ranggrupperna kan urskiljas.

[11] Här finns fog för en viss skepsis mot de skolor som betonar kontroll av kongruens hos undersärdragsvärden i fraser i analys. Uppenbart är kongruens avgörande för välformad generering. Men för analys visar rangbaserad chunkning hur liten betydelse eventuell kongruens behöver ha för parsning med syftet att identifiera led och deras sträckning.

Rangbestämning

Algoritmen innebär användning av den rangtilldelning som används i rangbaserad chunkning (Avh). Rangerna tillsammans med tolkningsregler utgör i sig en segmenterare (chunkare) och kan appliceras på ordklasstaggad text, eller med fördel som i schemaparsningen i klargjorda grammatiska fält.

Rangbestämningen som har presenterats i Avh och mycket koncist i (Wilhelmsson, 2008) innebär att varje löpord i analysområdet representeras med en fix rang. Systemet där varje löpord förses med en rang har konstruerats för att genomföra segmentering. Rangernas syfte är på ett sätt tvåfaldigt: att avgöra de olika chunksegmentens sträckning och att med tolkningsregler bestämma strukturtyp, huvudord och rektionshuvudord (för bl.a. PP).

Ordklass	Exempel	Rang
Som taggat som konjunktion	*Som målvakt var han bra.*	16
Preposition	*Till, för*	15
Ord i genitiv-kasus	*Kalles, bokens*	1 / 14
Determinator	*De, några*	5
Possessiv	*Dess, sitt*	4
Räkneord, grundtal	*43*	3
Adverb	*Ganska, bra, bort*	3
Particip	*Slående*	2
Adjektiv	*Grön, hoppfulla*	2
Räkneord, ordningstal	*43:e, första*	2
Måttsattribut (’mängdord’)	*Kopp, kilo, handfull, msk*	1,5
Persontitel	*Herr, cupvinnaren, tvåan*	1,5
Personligt pronomen	*Han, de*	1
Egennamn	*Karl, Karlsson, Paris*	1
Substantiv	*Idé, elefanter*	1

Tabell 4 Mappningen från ordklasstaggning till rang sker genom enbart ordklass (i de flesta fall), i taggning med särdragsvärden (löpord i genitiv) eller genom uttryckliga ordlistningar (för måttsattribut m.fl.).

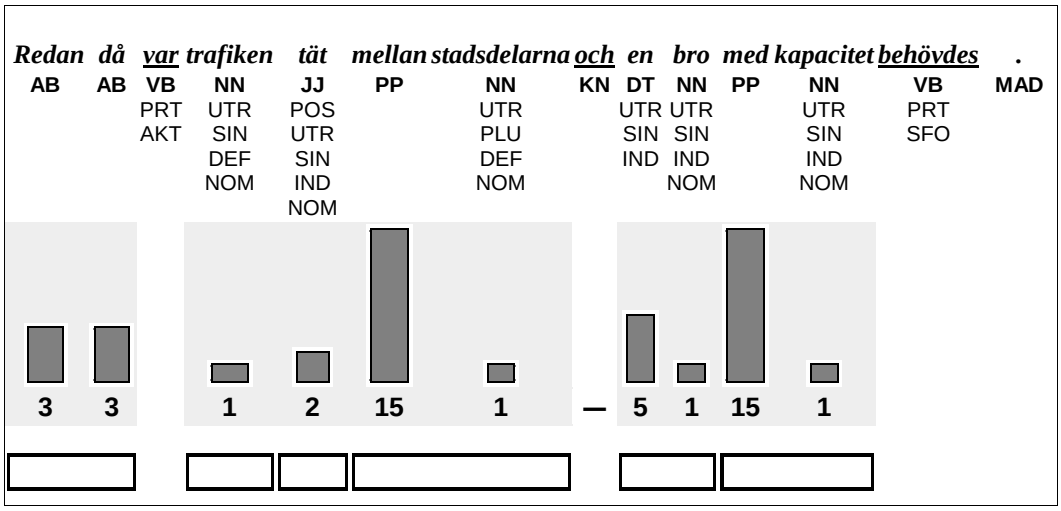
Det följande avsnittet bygger först på ett avsnitt i Avh, (Kap. 3) varifrån delar är inhämtade, och beskriver hur rangerna används för uppgifterna här.

Användningen av rangerna

16	15	14	5	4	3	2	1,5	1	**Chunk-typ**
Som	*i*	*Pers*			*ganska*	*hopp-fulla*		*plan*	*Som*-fras
Som						*nya*		*studenter*	*Som*-fras
	I		*några*		*bra*		*koppar*	*kaffe*	PP
				Sitt		*röda*		*hus*	NP
								Per Karlson	NP
					Ganska bra				Adverb
			De			*första*			NP

Tabell 5 Med hjälp av rangerna identifieras segment genom identifikation av en sekvens av fallande rang tills denna räcka bryts. Framförallt första och sista ordet i denna struktur har sedan betydelse för typbestämning av segment. Särdragsvärden (tagginformation utöver ordklass) används enbart i mycket liten omfattning.

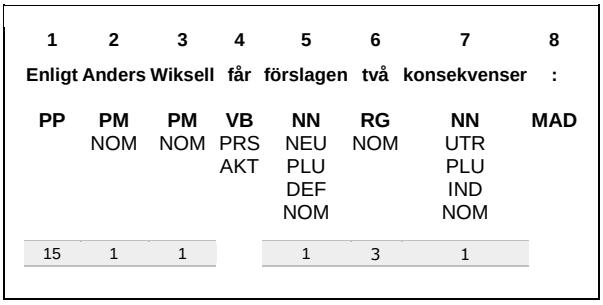
I nedanstående figur och beskrivningen härefter visas hur metoden används för att dels finna (avgränsa) strukturella enheter och senare klassificera dem. – Men i det aktuella projektet är alltså adverbialen givna och metoden används alltså enbart för att typbestämma dem samt avgöra huvud- och rektionshuvudord!



Figur 4 En illustration av ranger som stapellängd visar hur en högre stapel än den till vänster i ordsekvensen indikerar ny segmentstart (hb06a-011). De begränsade primära leden är understrukna.

Den algoritm som använder detta rangsystem i parsning kan beskrivas som nedan. [Fr Avh]

1. Områdena där de aktuella strukturerna ska identifieras behandlas först så att varje löpord med ordklass får en rang enligt ovan, som i Figur 5.



1	2	3	4	5	6	7	8
Enligt	**Anders**	**Wiksell**	**får**	**förslagen**	**två**	**konsekvenser**	**:**
PP	**PM** NOM	**PM** NOM	**VB** PRS AKT	**NN** NEU PLU DEF NOM	**RG** NOM	**NN** UTR PLU IND NOM	**MAD**
15	1	1		1	3	1	

Figur 5 Fundament och efterdel genomlöps och ranger sätts (ab02b-040).

2. En genomlöpning startar längst till vänster i varje sådant område (fält). Algoritmen innebär att varje nytt ord ska ha högre rang (lägre tal) eller lika rang som det föregående för att tolkas som del av samma segment (chunk).

   – Om rangen däremot är lägre (högre tal) än den till vänster så är tolkningen att föregående chunk avslutats och en ny chunk inleds med det aktuella ordet.

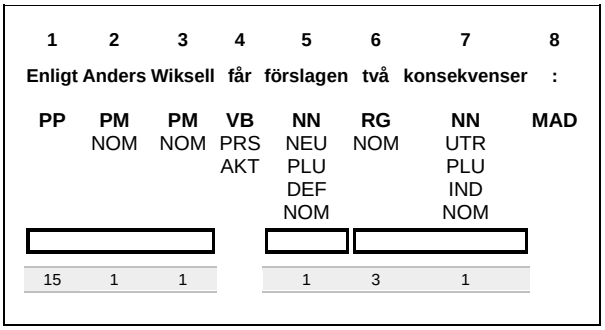
   – Detsamma gäller när två ord av rang 1 finns bredvid varandra – men där görs en speciell undersökning så att fulla personnamn *Bea (1) Karlsson (1)* räknas som samma struktur om båda är taggade som egennamn (*PM* i SUC) och det första ordet finns i en listning av förnamn och/eller det andra ordet finns i en listning av efternamn.

3. Efter denna genomlöpning kan frastypen fastställas.

   – Om det första ordet i ett segment är preposition (rang 15) är frasen en PP.

   – Om det första ordet i en chunk har rang 16 är strukturen *som*-fras.

   – I annat fall är strukturen något annat, företrädesvis nominalt: en NP, adverbfras, adjektivfras eller liknade, beroende på det sista ordet som är huvudord och avgör ’frastyp’. Chunktypen avgörs främst utgående från dess huvudord, som också specificerar om en NP kan fungera som adverbial.



Figur 6 Efter rangtilldelningen skapas inledande chunksegment.

I Figur 6 visas hur två nominala strukturer framträder efter finitet med den rangbaserade chunkningen – dessa är alltså subjekt och objekt i satsen. Följande specialfall finns dessutom.

- Samordnande konjunktioner i fälten (ej primära konjunktioner) innebär att den pågående chunken fortsätter oavsett föregående och efterföljande ord.[12]

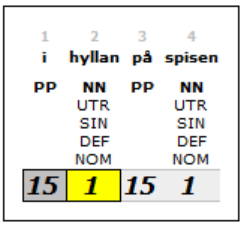
- Ord i genitiv har dubbel rang: 1 och 14. Denna rangsiffra är satt som en analogi till en kortlek (där en sekvens av fallande valörer bildar ett segment) där genitiv är ’ess’. Ord i genitiv fungerar därmed liknande konjunktioner och låter generellt chunken fortsätta det föregående (rang 1) och inbegripa det följande (14) såvida kommande chunk inte är preposition (rang 15) eller konjunktionen *så/såsom* (rang 16) i den sekvens av stadigt sjunkande rangsiffror som bildar en chunk.

Om huvudorden och rektionshuvudorden för den aktuella uppgiften

Eftersom det som är indata till det aktuella projektet är fulla adverbial med eventuella efterställda attribut är inte huvudordet (eller rektionshuvudordet för PP/*som*-fraser) detsamma som det sista löpordet utan kan ofta beskrivas (med en beskrivande term hämtad från matematiken) som *det första lokala minimumet*[13]enligt rangerna (ett eventuellt efterställt attribut bryter som regel av sekvensen av nedåtgående rangtal). Nedanstående bilder från prototypen visar hur det lokala minimumet så att säga innebär den sista positionen (ordet) av en sekvens av nedåtgående rangtal innan nästa rang blir ett högre tal eller liknande.

[12] Detta gäller inte alla konjunktioner, enligt hur denna klass ser ut i SUC: Parvisa samordnare (t.ex. *både/och, varken/eller*) och vissa andra (t.ex. *så*) fungerar inte på detta sätt.
[13] Om ett lägre tal ses som en högre rang kunde *lokalt maximum* vara ett lika beskrivande namn.

Bilderna är hämtade från den aktuella webb-prototypen. Korrektheten för dessa viktiga steg är inte hundraprocentig.



Figur 7 PP: En sekvens som inleds med preposition (rang 15) är PP. Det lokala minimumet blir rektionshuvudord.

Ett fall av huvudordssekvenser uppvisar mer av en teoretisk oklarhet, och är de fall av framförställda (eller enligt den alternativa tolkningen, efterställda) substantivbestämningarna som *i en liter mjölk*. Genomgående för denna ansats är att den rangbaserade chunkningen beskriver mängdordet (*liter*) som framförställt attribut och det påföljande som huvudord (*mjölk*). Här finns dock andra synsätt i litteraturen som hos Ljung & Ohlander (1971) där istället det påföljande ordet (*mjölk* i exemplet) kallas innehållsattribut.

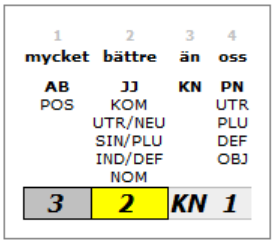
Den ovanstående användningen av rangsystemet levererar information om frastyp, huvudord etc. vilket alltså vore omöjligt att inte känna till i frågeledsbestämningen.

6 Vidare steg med icke PP-formade adverbial

För icke-PP-formade adverbialförekomster gäller för det första att de kan vara specialfall såsom satsadverbial och fasta utryck med särskilda omfrågningsegenskaper. För de övriga gäller dock att det frastypsbestämmande huvudordet (eventuellt i grundform) i allmänhet utgör tillräcklig mappningsnyckel och att mappningen för adverbial med strukturerna AdvP, AdjP, NP, PartP och bisats i hög grad är ’uppräkningsbara’. (Det betyder inte att den aktuella prototypen alltid gör rätt val, men perfekt korrekthet antas i princip kunna uppnås för dessa led.)

Adverbfraser

Adverbfraser har i detta projekt inte studerats speciellt. Mappningen är oftast okomplicerad: *därborta: var*. Det är en gruppering med många satsadverbial och liknande (t.ex. *fortfarande*), vilka saknar frågeled.



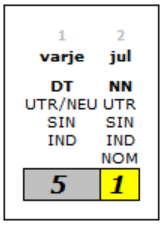
Figur 8 AdvP: Huvudordet som det sista ordet före en (konjunktionsinledd) jämförelsedel.

Adjektivfraser och participfraser

Som participfraser räknas sådana där huvudordet enligt taggingen i SUC 2.0 är particip (PC). Gruppen sönderfaller i presens- *(förvånande, halvspringande)* och perfektparticip *(datorstyrd, kokta)*. När dessa förekommer som huvudord i fraser blir de som grupp närmast omfrågade med *hur*. En närliggande omskrivning som ger liknande funktion (medel/sätt) är *på vilket sätt/vis*.

Nominalfraser

Nominalfraser som adverbial är en intressant företeelse. Gemensamt för alla förekomster av NP-adverbial är att de kan fungera som subjekt eller objekt/predikativ när syntaxen kräver sådan tolkning. Det är alltså en inte helt trivial uppgift i parsningen att känna igen dem som just adverbial i det aktuella läget. (I prototypgränssnittet där användaren skriver in led är dock tolkningen genomgående att det som kommer som indata är adverbial.) Majoriteteten av de NP-formade adverbial som påträffas i svensk text står för tidsuttryck, vilka motsvaras av *när* (en punkt i tiden) eller en duration; *hur länge*. De NP-formade tidsadverbialen är relativt vanliga och förekommer i mängder av olika former (inkl. tidpunkter i en nära obegränsad mängd).



Figur 9 NP: Strukturtypen har avgjorts genom att inledningsordet inte är preposition, *som*-konjunktion eller bisats. Det är i detta läge det lokala minimumet (här valt eftersom inget följer) som avgör strukturtyp (här NP). Hade ledet istället haft particip, räkneord, pronomen eller adjektiv som lokalt minimum skulle det därmed tolkats som PartP, ”räkneordsfras” (en sorts NP), pronomenfras (NP) respektive AdjP.

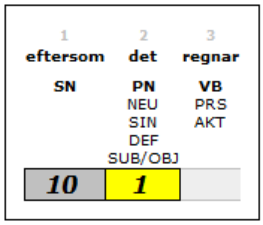
Ovanstående figur illustrerar även den mer generella frågeställningen om vad som är rätt frågemotsvarighet (speciellt i ett läge utan kontext). *När* verkar ibland gångbart men en ’frekvenstolkning’ kan här även ge *hur ofta*.

När	*Var/var(å)t*	*När/var*	*Hur (mycket)*	Ej omfrågningsbara led inkl. satsadv.
Tis *En timme* *1912* *Ett ögonblick* *Två minuter* *Jan-mar* *Den 27 augusti 2005* *Ett slag*	*Ett stycke* *Ett stenkast* *De första metrarna* *Två trappor upp*	*Första varvet*	*En smula* *Lite grand* *En aning*	*Tack och lov* *Min själ*

Tabell 6 Några möjliga frågemotsvarigheter för NP-adverbial.

Bisatser

Bisatsadverbial som beaktas har adverbialfunktion (de med nominal funktion omfrågas i allmänhet med *vad*). De är generellt ordklasstaggade som subjunktion (SN). Ett fåtal, bl.a. *som*-satser är taggade som frågande/relativt adverbial (taggen *HA* i SUC). Grupperingen innehåller huvudsakligen medlemmar med klara mappningar: *eftersom – varför*.



Figur 10 Bisats: Strukturtypen bisats känns igen genom (rangen/ordklassen för) inledningsordet. Det som här markerats som lokalt minimum i gult därefter har inte behövts beaktas hittills för bisatser.

Bisatser: Normalt adverbiella led som objekt

Ett faktum som kan vara lätt förvirrande är att ett flertal bisatstyper som inleds med en subjunktion med ”adverbialkaraktär” såsom *hur*, *var*, *hurdan*, *varför* omfrågas som nominala led (han undrade *var….: vad)*. Det första som ska konstateras är att vissa bisatsled med adverbialanknuten semantik som t ex *hur*-bisatser generellt är nominala.

Hur det ska gå (NOM) vet vi ej. → *Vad*

Undantaget därifrån är generaliserade bisatser:

Hur det än går (ADV) cyklar vi → *Ø, ”under vilka förutsättningar”*
Var vi än går (ADV) regnar det → *Ø, var*

Somliga led kan både ha *adverbial* och *nominal funktion*.

När vi cyklar (ADV) regnar det → *När*
Hon undrar när vi cyklar (ADV) → *Vad*

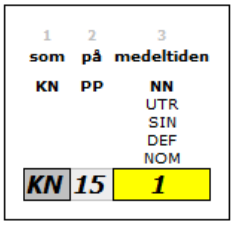
Som redan klargjorts är det inte mappningsdelens uppgift att avgöra om ledet verkligen är adverbiellt.

Som-fraser

Som-fraser kallas här, i enlighet med Mambans (Teleman, 1974) terminologi, de fraser som inleds med *som* taggat som konjunktion (KN) i SUC 2.0. [14]

[14] *Som* förekommer i SUC 2.0 med tre olika ordklasstaggar: konjunktion (KN), frågande/relativt pronomen (HP), samt frågande/relativt adverb (HA). KN-varianten som här beaktas inleder ej bisats. HP-varianten inleder efterställda relativbisatser *(som jag känner)*. HA-varianten inleder *som*-satser *(som det verkar)*, alltså ett bisatsadverbial.

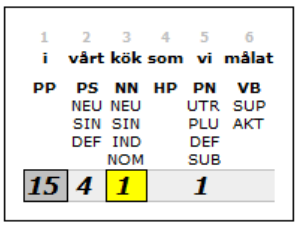
Hanteringen av *som*-fraser liknar den för prepositionsobjekt. Hela ledet *(som målvakt, som målvakten, som i Spanien)* kan närmast omfrågas med *hur*. Alternativet är emellertid att ordet *som* plus *som*-frasens rektion, vare sig den är nominal eller adverbiell: *som vad, som vem, som var, som när, som på vad* etc. I princip är hanteringen av *som*-fraser därför den allra mest komplexa – *som*-frasens rektion har i sig formen av något annat obegränsat led, vilket måste analyseras.



Figur 11 *Som*-fras: *Som*-fraser kan kännas igen antingen med inldeningsordets ord-tagg-kombination, eller den utsatta rangen 16 (ej i bilden här), och har antingen ett adverbiellt eller nominalt komplement. När komplementet som här är PP blir det lokala minimumet rektionens (rektions) huvudord. Både komplementshuvudordet *på* och dess huvudord är egentligen nödvändig data för att avgöra motsvarande frågeled (jämför *som på hyllan*).

7 Vidare steg med PP-formade adverbial

PP-adverbial känns igen enkelt genom inledande preposition (rang 15). De har ett mycket stort antal olika fall av olika frågemotsvarigheter som här försökts täckas in med regler. Ungefär en fjärdedel av prepositionerna är relativt frekventa och har unika möjligheter till olika mappningar beroende på rektionshuvudordet för en viss preposition eller ibland en grupp prepositioner som grovt sett fungerar på samma sätt. Regler som skrivs angår ofta preposition och rektionshuvudord eller dess grundform, vilken tas fram med en särskild grundformsfunktionalitet som byggts i projektet (se avsnittet i *Tekniskt utförande*).



Figur 12 PP: Huvudordet är här det första ordet, eftersom det är preposition (rang 15), och det lokala minimumet, *kök*, markerat i gult, är rektionshuvudord före den efterställda relativbisatsen.

En stor del av arbetet har ägnats åt försök att korrekt behandla PP-formade adverbial. Det finns mellan 150 och 200 prepositioner i svenska uppmärkta som sådana enligt riktlinjerna i SUC 2.0. En blick på hur prepositionsfraserna fördelar sig i antal ger grovt sett följande bild.

- En stor del av de olika prepositionerna utgörs av relativt ovanliga prepositioner som har ganska enkla mappningar till frågeled (*inunder* osv.). De är i många fall sådana där komplementsdelen inte behöver undersökas alls för att ge rätt frågeled. Dessa kan kallas *starka*.
- En mycket stor del av prepositionsförekomsterna i fri text inleds av en grupp prepositioner som kan kallas *svaga*. Dessa karakteriseras av:

    - att det är väldigt frekventa (*i, på, till*)
    - att de på olika sätt har möjligheter att mappas till olika frågeled, bl.a. beroende på rektionshuvudord.

    De olika svaga prepositionerna kan dra nytta av en gemensam kategorisering av rektionshuvudord – t.ex. är grupperingar av tidsangivande substantiv som *år, dag* osv. ofta användbar, då dessa ger *när* för bl.a. *i*.[15]

Prepositioner som *bland, efter, från, för, före, genom, i, inemot, med, mellan, mot, om, på, till, under, vid* inleder på ett tydligt sätt *flera olika* frågeledslag från ett frågeperspektiv och måste behandlas olika beroende på deras komplementdelar. Dessa prepositioner kräver en speciell hantering. Kanske är *i* den preposition som har flest undantagsfall och har inneburit mest regelskrivande. Eftersom de nämnda PP-slagen är så frekventa och samtidigt svarar mot så många olika fall har de utgjort en huvuddel i regelskrivandet för att uppnå viss korrekthetsnivå. Under detta korta projekt har grunden för fortsatt arbete lagts. Det finns nu teknik för att fortsätta att förbättra den aktuella mappningen och möjlighet att lägga till nya regler. Det ska dock konstateras att det i aktuell version förekommer en hel del fel även av det enklare slaget. Detta beskrivs vidare nedan. Först ska här beskrivas två specialfall av PP som särbehandlas: agentadverbial och s.k. prepositionsobjekt, vilka båda renderar pied piping-form som frågeled. Dessa föranleder analys av rektionen, vilken ofta har substantivhuvudord, detta tas upp därefter.

[15] Med vissa attribut och former som *i fem år* (jämför med *i år*) är tolkningen duration (*hur länge*). En preposition som *om* tar å andra sidan en period (duration) som komplement och ger *när*.

Agentadverbial

Agent-adverbial känns igen genom att de strukturellt är *av*-PP och har ett tillhörande (här: i gränssnittet angivet) huvudverb i passiv diates. Grupperingen omfrågas med pied piping eller rektionsframflyttning plus strandad (”isolerad”) preposition. Ett led motsvarande den *nominala* rektionen måste därmed tas fram.

Av-PP är inte helt enkla att hantera pga. att rektionsdelen måste undersökas för att ger *av vad, av vem* eller *av vilka* – men i praktiken även t.ex. *vilket företag*. Det empiriska arbetet har tydliggjort att frågeformen *av vilka* företrädesvis svarar mot plural med mänskliga referenter. T.ex. *av en samling skivor* ger t.ex. snarare *av vad*.

Prepositionsobjekt

De led som ofta har kallats prepositionsobjekt eller liknande i litteraturen, men som speciellt av Hellberg (2003) ändå har motiverats som adverbial har speciella regler för omfrågande. I Avh konstaterades att ett sammanfattande *hv*-frågeord saknas eller att den nominala rektionen är vad som efterfrågas. *Jag lyssnar på musik* saknar alltså, (eller förekommer sällan), med något i stil med ’*Vadpå’ lyssnar jag*. Istället förekommer rektionsframflyttning med strandad preposition, *Vad lyssnar jag på*, eller pied piping *På vad lyssnar jag*. Adverbialtypen prepositionsobjekt är ett av de tydligaste exemplen på att uppgiften adverbialkarakteristik från ett frågeledsperspektiv kräver extra lexikala resurser för identifikation, i detta fall verbvalensresurser.

I det praktiska arbetet med mappningen skrivs regler för PP-adverbial utan hänsyn till deras huvudverb och dessa ges därmed en mappning som de vid användning av programmet tilldelas, förutsatt att de inte fungerar som prepositionsobjekt. En eventuell sådan matchning med ett huvudverb avgör alltså vid den praktiska frågeledstilldelningen. Det intressanta är här att redan prepositionsfrasen i sig, ibland vid regelskrivandet inte kan tänkas ha någon annan roll än som prepositionsobjekt, och detta kan sägas utan kännedom om aktuellt huvudverb. *På musik* ovan är sådant exempel som inte rimligen kan ge någon annan motsvarighet än *på vad*.

Några exempel på olika PP-adverbials egenskaper

För många PP-formade adverbial är fallet att de inte så enkelt låter sig omfrågas i svenska. En *frånsett*-PP kanske omfrågas om alls med pied piping, ’frånsett vad/vem’.

Under-adverbial förekommer en majoritet av gångerna med temporal betydelse (*när*) och *var*-betydelsen är i testtexter ett undantag tillsammans med fixerade konstruktioner (t.ex. *under hot → hur*). I de första versionerna av frågegenerering fanns en mappning som kunde sägas härröra från en sådan prototypisk uppfattning av en prepositions semantik snarare än en frekvensgrundad – ett talande exempel är *till*-adverbial vilka tidigare kunde ges *vart* som grundtolkning. Detta utgjorde en betydande felkälla i de första versionerna.

Till-PP exemplifierar en grupp av mycket vanliga PP-adverbial som här kallas *svaga* och där komplementet har mycket stor betydelse för semantiken och frågemotsvarigheten. [16] När det gäller *till*-adverbial är den vanligast förekommande motsvarigheten kanske inte *var/varåt* eller *till(s) när*, utan just pied piping-frågan. Detta är möjligt att skönja i exemplet nedan även utan att känna till satsers huvudverb.

Det nedanstående är en uppställning över faktiskt förekommande *till*-adverbial med frågeledförslag.

Till en början	*När/hur*	
Till en viss del	*-/Hur*	
Till exempel	*-*	
Till exempelvis Hannes rågbröd (s. 231)	*till vad*	*dvs. pied p.*
Till flygande fanor och klingande spel	*hur/till vad*	*ev. pied p*
Till för några decennier sedan , då det ännu var tillåtet ,	*Till(s) när*	
Till glädjestunderna	*till vad*	*dvs. pied p.*
Till hösten	*När*	
Till kollektionen	*till vad*	*dvs. pied p*
Till kongressen	*Vart/till vad*	*ev. pied p*
Till lokalerna längst ned	*vart*	
Till läkarmottagningen	*vart*	*ev. pied p*
Till ny 1:e vice ordförande	*till vad*	*dvs. pied p*
Till pluskontot	*till vad/vart*	*dvs. pied p*

Exemplet *till*-adverbial visar tydligt på behovet av valenslexikon för att avgöra speciellt de fall som har både en möjlig prepositionsobjektsroll och även annan roll. I implementationen används databasen till NEO (1995–96) som i Avh: om en inledande preposition förekommer i valensdata för ett tillhörande huvudverb blir tolkningen prepositionsobjekt, vid en sådan matchning väljs alltså pied piping.

[16] Här har frågan om huruvida frågeledet *var/varåt* är ett och samma i svenska aktualiserats (det verkar finnas etymologiskt stöd för det men uppenbar utbytbarhet saknas i normalsvensk användning).

Pied piping

Termen pied piping,[17] som verkar sakna en enkel svensk översättning, innebär att ett PP-format adverbialled omfrågas med prepositionen plus ett *hv*-led motsvarande dess komplementsdel; *på taket: på vad, under denna tid: under vad*. Pied piping-formen och dess systerform är viktiga begrepp i detta arbete.

Den teoretiska frågeställningen om huruvida detta är en fråga som gäller hela ledet eller egentligen snarare komplementsdelen kommer att lämnas därhän. Det är oavsett hållning nödvändigt att identifiera dessa led, för att antingen omfråga dem på nämnda sätt eller utesluta dem från frågegenerering. I detta projekt har syftet varit att generera förslag till pied piping-versioner av dessa led. Det finns sammantaget många PP-led – de nämnda agentadverbialen, prepositionsobjekten och även en del andra led – där ingen annan lösning känns rimlig; exempelvis *om Spanien*. För ett sådant exempel blir det rimligt (dessutom helt utan kännedom om huvudverbet; men med antagandet att det måste röra sig om valenskoppling som gör adverbialet till ett prepositionsobjekt) att säga att just pied piping är den bäst anpassade frågeformen, om någon.

Möjligheten att producera pied piping-lösningar för PP-formade adverbial är något som öppnar för en del frågor. Som nämnts finns PP-typer vilka måste omfrågas så om de ska kunna omfrågas alls: detta gäller t.ex. *av*-PP. Men i många fall är pied piping-varianten en grammatiskt godtagbar men funktionellt sett mindre lyckad frågeform: det verkar som om de flesta PP kan omfrågas med pied piping, men en fråga som *På vad sitter Kalle/Vad sitter Kalle på* verkar innebära att användaren av ett frågesystem har kännedom om att ’Kalle sitter *på* något’ – till skillnad från *var*.[18] Detta gör frågetypen mindre generell och frågan svårare att hitta i t.ex. ett frågesystem som producerar en stor mängd frågor. När det är möjligt föredras enkla *hv*-led som korrekta resulterande frågeled i detta projekt.

I takt med att de statliga verken bolagiseras, dvs. *i takt med*-adverbial är typiska fall som är undantag från grundregeln (för *i* är *var* annars den vanligaste motsvarigheten; satsadverbial och temporala tolkningar är vanliga undantag). När skulle detta led fungera bäst som svar på en fråga, och för vilken frågetyp? Fallet exemplifierar en semantisk kategori *hastighet* som ibland kan omfrågas

[17] Namnet *pied piping* kommer från bröderna Grimms folksaga *Der Rattenfänger von Hameln (The Pied Piper of Hamelin)* och beskriver hur prepositionskomplement (rektion) tas med av prepositionen vid spetsställning, namngivet av John R. Ross (1967); [från Avh].

[18] Systerkonstruktionen *Vad sitter Kalle på* innebär likaledes en eventuell kännedom om en nominal PP-rektion.

med *hur snabbt*. Den aktuella konstruktionen har dock en annan särskild betydelse och det är möjligt att säga att den ’lånar’ hastighetsaspekten.

Det ligger här nära till hands att använda en pied piping-konstruktion, i detta fall rörande komplementet till *i takt med*. Det resulterar för detta fall i frågetypen *I takt med vad?* Pied piping frågor har som poängterats fördelen att inte vara felaktiga i sak men att de generellt placerar komplementsdelen i fokus.

Som beskrivits utgör pied piping i många fall en antaget mindre användbar men säker lösning, som kan sägas gälla rektionen. Det kan bero antingen på att prepositionstypen, t.ex. *utöver*-PP, inte har någon särskilt mycket bättre motsvarighet över huvud taget. Exempel: *enligt: enligt vem/vad, frånsett: frånsett vad/vem, tvärtemot: tvärtemot vad/vem*. I vissa fall (för många PP-adverbial, prepositionsobjekt och agentadverbial och för specifika förekomster; *på musik*) den enda möjliga lösningen. Det är tyvärr inte så att pied piping för alla PP-adverbial innebär en alltid tillgänglig enkel nödlösning. Att PP-adverbial kan omfrågas genom en konstruktion med ’preposition + *hv*-motsvarighet’ för komplementsdelen som oftast är ett nominalt led och motsvarigheten till detta nominala led måste alltså avgöras. Att ge frågemotsvarighet för vanliga nominala led är emellertid inte enbart ett val mellan *vad/vem/vilka*.

Pied piping: nominalkarakteristik som en deluppgift

Medan adverbialen har många olika motsvarande frågeled har de nominala leden, som det har verkat, inte fullt så komplicerade förutsättningar. *Vad* kan ses som default-värde och motsvarar nominala satsförkortningar inklusive infinitivfraser och bisatser samt många NP med icke-animat referent. *Vem* kräver en animat referens och *vilket/vilken/vilka* kräver i allmänhet en begränsad mängd alternativ och/eller är vad som skulle kunna kallas attribut-frågor (*vilken häst*) – därmed svarar frågeledet inte mot ett helt adverbial och faller utanför detta projekts råmärken.

En lärdom är från det praktiska arbetet är att animathet här får betydelse: När rektionshuvudordet står i plural tenderar frågeledet att vara *av vilka* endast om rektionshuvudordet är mänskligt (faktiska exempel: *av de människor som är aktiva inom musiken, av de tre studerande, av barn och dårar*). Om rektionshuvudordet är inanimat plural (*av de många beröringspunkterna*) är frågeledet med fördel *av vad*.

Främst på grund av adverbialtyperna *som*-fras, prepositionsobjekt, men även eftersom andra adverbial kan eller måste efterfrågas med pied piping *(På vad lyssnar…)* eller med isolerad preposition *(Vad lyssnar… på)* krävs ett

klargörande om *vem* eller *vad*. Att avgöra detta frågeled för godtyckligt nominalt led (eftersom godtyckligt nominalt led här kan vara rektion) gör att uppgiften adverbialkarakteristik egentligen inbegriper uppgiften nominalkarakteristik.
Pied piping är ett av de tillfällen då uppgiften kommer i direktkontakt med substantiv – PP-rektionens huvudord är ofta ett substantiv. Denna vidsträckta lexikala mängd gör att ambitionen om fullständig täckning får anses mindre plausibel, för nominala led, och därmed också för adverbiella.

Om frågan gäller ett nominalt led, som subjektet i *Volvo köpte hamnen*, så är knappast *Volvo* bäst representerat med *vad* eller *vem* utan förmodligen med *Vilket företag* eller liknande.

Mänsklig referent

För att avgöra frågeled för t.ex. prepositionsobjekt där pied piping allmänt används måste i princip en funktionalitet för nominala led, dvs. PP-rektioner osv. fungera. Bestämning av mänsklig referent är här relevant. Det är ett avgörande som får betydelse för att skapa frågor som *vem* (för nominala led) och *till vem, som vem* och *som mot vem* (för adverbiella).

Personrefererande (vem)		Ej personrefererande (vad)	Oklara fall (vilka)
Jag	Han	Det	de
mig	honom	den	dessa
mej	vi	detta	dom
dig	oss		dem
dej	ni		
henne	er		
hon			

Tabell 7 En trivial uppdelning av pronomen utgående från motsvarande frågeord gör att de fall som kanske är oklara, till höger, kan omfrågas på med vilka och därmed undslippa klargörandet.[19]

I arbetet med schemaparsern för svenska fanns en rudimentär sådan funktionalitet som över huvud taget gav ledtrådar om animathet, dvs. även för andra varelser som djur och även organisationer med förmåga att handla ’som människor’.[20] Mänsklig referent innebär dessutom den mest prominenta NP-typen, med konsekvensen att den är benägen att besätta subjektsrollen. I det sammanhanget är dock pronomen med mänsklig referent olika – de i objektskasus är naturligtvis inte subjektsbenägna. I den aktuella uppgiften är

[19] Den aktuella diskussionen om ett politiskt korrekt *hen* innebär troligen ännu en medlem i *vem*-kategorin.
[20] I somliga lexikon som t.ex. *Lexin – Svenska ord* (1998) finns animathet hos subjekt kodat med A/B för mänskliga samt X/Y för icke-animata referenter.

dock också ett pronomen i objektkasus sådant att det omfrågas enligt ovan och kommer att medtas.

För pronomen-märkta ord (*PN* i SUCs taggset) gäller att en grupp har mänskliga eller liknande referenter, en har inanimata och somliga är oklara vad beträffar referentens beskaffenhet.

Här finns somliga ovanliga undantag att beakta men tabellen ovan visar de huvudsakliga tendenserna.

Egennamn

För att avgöra egennamns (*PM* i SUC) referenter krävs uttryckliga listor. Det antas vara ett ofrånkomligt resurskrav. I detta fall har docent Dimitrios Kokkinakis varit behjälplig både med för- och efternamn, samt organisations- och platsnamn. Listorna på för- och efternamn har därefter förstärkts genom SUC 2.0.

Förnamn	Efternamn	Platsnamn	Organisationsnamn
Ca 12480 ingångar	Ca 9400 ingångar	Ca 6690 ingångar	Ca 2460 ingångar

Potentiella svårigheter uppkommer bl.a. när efternamn är platsnamn. Andra grupper av egennamnsuppmärkta ord är förkortningar som naturligtvis kan representera organisationer och ämnen, dvs. både sådana som kan fungera som människor och som ting. Dessa omfrågas dock oftast inte med *vem*. Organisationer är inte helt uppenbara från omfrågningssynpunkt, *'från Volvo/från Kulturrådet'* omfrågas kanske enklare med *från vilket företag/från vilken instans* än med *från vad/vilka*. Någon sådan beskrivande underkategori finns för närvarande inte med för organisationsnamnen.

Substantiv

Huvudord i form av substantiv (*NN* i SUC) har mänsklig eller animat referens. För denna mycket rika ordklass finns information implicit i den lexikala resursen *SweFN*. Det gäller de semantiska kategorier till vilka de lexikala enheterna i resursen är kopplade. Ca 470 av dessa semantiska kategorier har medlemmar som kan ses som substantiv (här stämdes ingångarna av mot den med SALDO utformade grundformsfunktionaliteten). Somliga av dessa kategorier innehåller snarare particip med möjliga substantivtolkningar.

Kategorinamnen (som *person_by_vocation* eller *Furniture*) ger tydliga ledtrådar för insamlandet av personrefererande ord resp. andra substantiv. Vissa andra kategorier, t.ex. *Touring*, har både *turist* och *turism* som medlemmar och är därmed inte lika användbara för att snabbt klassificera en medlem.

Substantiv med antagen mänsklig referent utgående från tolkning av kategorinamn i SweFN
Ca 1330 ingångar

Detta arbete med PP-rektioners semantik för omfrågning är överförbart till arbetet med vanliga nominala led. Det betyder att detta arbete härmed utökas till att gälla i princip samtliga hela obegränsade funktionella satsled: utöver adverbialleden också subjekt, egentligt subjekt och objekt/predikativ. Användningen av semantisk information från SweFN är för närvarande begränsad i implementationen bl.a. av ovanstående skäl. Se vidare i Appendix.

Svaga prepositioner

För frågegenerering är PP en gruppering som av ovanstående och nedanstående anledningar är en avskräckande komplicerad adverbialtyp i sammanhanget. Som nämnts finns flera undantag att ta i beaktande när det gäller frågeledsbestämning. Dessa rör t.ex. allmänna svårligen omfrågade undantag (ett axplock: *i huvudsak, i konkurs, i kö, i styrka, i timmen, i särklass, på tal, på tomgång*), ovan nämnda PP-adverbial av typen prepositionsobjekt, agentadverbial och satsadverbial. Prepositionsobjekt har i tidigare arbete visat sig kunna identifieras ganska väl genom verbvalensinformation från databasen till NEO (1995–96), vilket används i implementationen. De PP-formade satsadverbialen är delvis även de av obegränsad längd, enligt de exempel som ges i t.ex. SAG (*enligt …s mening* etc.) och är därmed inte hanterbara enbart genom enkla listningar, men kan oftast kännas igen, liksom agentadverbial.

Med dessa undantag avskalade här återstår den stora utmaningen i de PP-formade adverbial som är vanliga innehållsadverbial av olika slag. Dessa adverbial kan uppdelas i sådana där prepositionen har en tungt vägande roll för det frågeled som motsvarar ledet. *Bortifrån, inuti* m.fl. tycks ha stor betydelse för vilket eller vilka frågeord som ledet motsvarar. Dessa prepositioner kommer här att benämnas *starka*. En annan grupp av prepositioner som är mycket vanliga är den med *i, på, med*. Dessa kommer här att benämnas *svaga* eftersom rektionens huvudord har avgörande roll för ledets omfrågningsaspekter. Även om denna uppdelning är en förenkling och saknar skarpa gränser antas den kunna tjäna sitt syfte att visa den stora skillnaden mellan olika prepositioners funktionssätt.

För grupperingen svaga prepositioner är situationen att det är en relativt stor möjlig frågevariation (*på medeltiden, på mitt sätt, på morgonen, på Irland*). Många prepositioner faller mittemellan starka och svaga prepositioner, t.ex. *mittemellan*, som åtminstone kan motsvaras av både *var* och *när*.

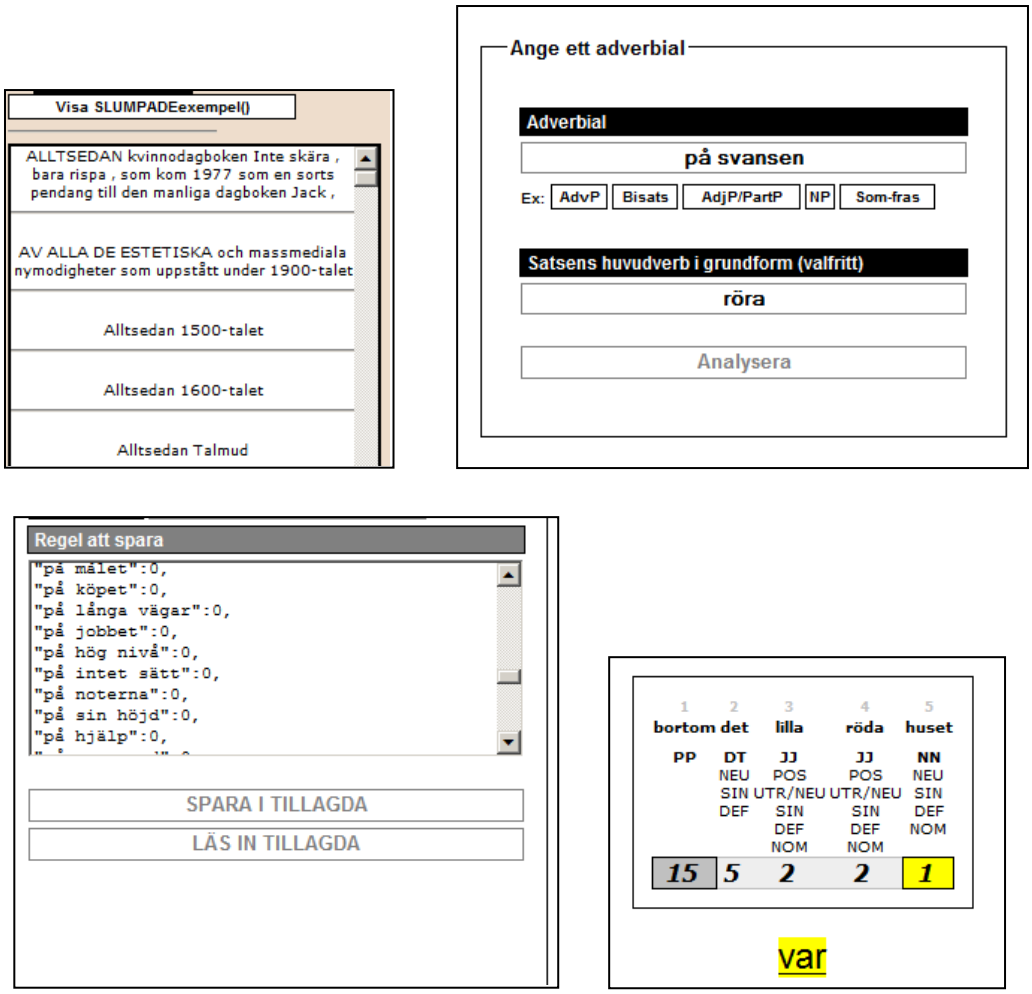
Detta pekar på att rektionshuvudordet, vilket dessvärre ofta tillhör den rika ordklassen substantiv måste undersökas för att ge ledning om den fråga, om någon, som ledet kan omfrågas med.

8 Om det tekniska utförandet och regelskrivning

Den aktuella versionen av programmet fungerar långt ifrån felfritt för uppgiften. Den tillhandahåller dock medel för att förbättra analysen genom att i många fall tillåta direkt uppdatering av reglerna, dvs. skriva nya överordnade undantagsregler, delvis direkt i gränssnittet (ej i versionen på Internet). Detta avsnitt tar upp några aspekter rörande det praktiska arbetet. Flera funktioner har skrivits för att underlätta det praktiska arbetet med regelskrivandet.

Huvudgränssnittet har som syfte att visa upp det resulterande frågeledet när ett adverbial skrivs in i ett formulär. Användargränssnittet är tänkt att åskådliggöra de delprocesser som leder fram till valet av frågeled: ordklasstaggning, ranger, strukturtyp, huvudord, rektionshuvudord/'lokalt maximum' och grundform. På samma gång är programmet en arbetsbänk för att i viss mån direkt formulera och prova nya regler.[21]

[21] Om gränssnittet körs i Windows-miljö är det särskilt enkelt att spara nya regler i kodfiler, eftersom det möjliggörs genom ett ActiveX-objekt. Detta rättighetsmässigt speciella steg lär även vara möjligt i (vissa implementationer av) HTML 5 samt i vissa andra miljöer.



Figur 13 Bilder från gränssnittet för regelförbättring. Uppe t.v. visas inhämtade huvudsatsadverbial från SUC som hämtats med hjälp av schemaparsern. Uppe t.h. visas formulär för inskrivande av adverbial för analys. Nedan t.v. syns formulär för direkt inskrivning, speciellt för undantagsfall. Nedan t.h. visas en analys och markering av huvudord samt rektionshuvudord.

Grundformslexikon

En användbar resurs för att skapa ett grundformslexikon för svensk text idag är SALDO (Borin, Forsberg, & Lönngren, 2008). Denna stora resurs finns i ett XML-format och är stor. En möjlig lösning vore enskilda webbanrop mot en databas under körning, vilket har undersökts. Den idé som här används innebär att funktionaliteten för substantiv lagras som JavaScript på ett relativt platssnålt sätt, det har fått namnet Kort-SALDO. Kort-SALDO är ett paket som tar fram grundformen med en klientbaserad JavaScript-lösning utgående från SALDO. Syftet med det är att åstadkomma grundformsfunktionaliteten så platssnålt som möjligt. Metoden går ut på att enbart spara grundformerna och ett förändringslexikon. Grundformsfunktionaliteten är byggd genom att koppla

möjliga tekniska suffix till det som behöver läggas till för att nå grundformen. När ett ord analyseras tas alla tekniska suffix för ordet fram och de möjliga förändringar av 'suffixet' som för någon ordform lett till rätt grundform kombineras med respektive prefix och bildar en samling strängar som testas mot ett grundformslexikon. På grund av de många möjliga lösningar som ibland uppstår finns ett undantagslexikon.

Processbeskrivning i slutprogrammet

Det följande är en beskrivning av programmet för frågeledsbestämning. Huvudstegen för frågeledsavgörande är i överensstämmelse med den schematiska beskrivningen som gjorts i den aktuella projektbeskrivningen. Indata till programmet är ett av användaren givet adverbial samt optionellt ett huvudverb för satsen där adverbialet förekommer.

Strukturbestämning: Adverbialet som införts analyseras strukturellt eftersom analysen företas på olika sätt beroende på struktur.

a. Ordklasstaggning. Ordklasstaggning sker med en ordklasstaggare som skapades i arbetet med schemaparsern. Den bygger på ordklasstaggningen i SUC 2.0 och ger för närvarande en utdata i HTML-format.

b. Rangtilldelning. Utgående från löporden och ordklasstaggningen tilldelas löporden ranger.

c. Regler för strukturbestämning avgör vilken strukturtyp adverbialet har.

Identifikation av de avgörande löporden.

a. I undantag av typen satsadverbial kan ibland hela adverbialet behöva matchas uttryckligen.

b. För *som*-fraser och PP-formade adverbial krävs att minst prepositioner och ofta rektionshuvudord i grundform tas fram.

c. För övriga typer som NP, AdvP, PartP, AdjP och bisatser är huvudordet ensamt avgörande.

d. För agentadverbial gäller matchning av det angivna verbets grundforms valensdata (prepositioner) med den aktuella prepositionen, samt att verbet förekommer i grundform.

Val av frågeled utgående från avgörande löpord och olika skapade och extraherade resurser.

Praktiskt regelskrivande: konsekvenser av den exempelstyrda ansatsen

I det praktiska tillvägagångssättet används för PP-adverbial, vilket generellt är den mest krävande strukturkategorin, något som bäst kan beskrivas som basfall med undantag i programkoden. Ett exempel är kategorin *i*-PP som adverbial. Den är mycket frekvent och sönderfaller i många olika unika fall. Reglerna för denna typ av kategori skrivs företrädesvis på grundval av:

- matchning av hela adverbialet, för att hantera vissa konstruktioner som *i motsats till*
- matchning av delar av frasen, oftast prepositionen plus rektionshuvudordet som *i […] översättning*
- matchning av grammatisk information hos rektionshuvudordet, som substantiv, singular, obestämd form en undantagsform som matchar *i sak* m.fl. och som bildar en kategori som själv har undantag (t ex *i översättning*). Denna kategori ger hur som default-led men i översättning utgör ett undantag och innebär här *'inget'* (dvs. ej omfrågbar enligt reglerna).[22]

Undantagen antar olika strukturformer, de har samlats in och skrivits in i systematisk anda men de är ibland speciella regler som skapas direkt i gränssnittet. När ett adverbialexempel av PP-form visas upp är det i många fall möjligt att direkt i gränssnittet skapa regler som angår prepositionen samt rektionshuvudordet, eller dess grundform.

En svårighet med halvmanuellt regelskrivande är som nämnts att på förhand veta hur fallen fördelar sig i antal. Det är på förhand oklart när det lönar sig att beskriva en undergruppering som basfall med undantag därför att syftet är hög korrekthet och denna grundar sig på frekvens hos olika grupperingar och detta är väldigt svårt att känna till. Konsekvensen blir ett regelskrivande som i stora stycken tenderar att få en *ad hoc*-karaktär. Programmeringskoden saknar dock

[22] Denna bild av hanteringen av i-PP behöver kompletteras med de undantagsfall då verbets valensinformation ger en pied piping-fråga.

den elegans och struktur som man finner i implementationer för modellbygge av begränsade språkfragment. Det går att likna regelförändring genom genomgång av exempel med vissa typer av maskininlärningsbaserade metoder som *transformationsbaserad maskininlärning*, men det manuella tillvägagångssättet innebär trots sin bristande överblick över de fall som ska täckas ändå en del regelskrivning som får hjälp av programmerarens språkkänsla och idéer om generaliseringar på rätt ställen. I programkoden smälter mer generell hantering av adverbialtyper samman med undantagshanteringen, detta är något av en konsekvens av arbetssättet.

Kodstrukturen för avgörandet av frågeled för en viss adverbialtyp t ex *av*-PP kan alltså beskrivas på en abstrakt nivå som grundfall och lager av undantag. I praktiken blir denna programmeringskod ganska rörig och präglad av de många undantagen och de olika specialfall som använder matchning av huvudord, rektionshuvudord, ingående ordsegment m.m. för att avgöra frågeled. Detta ger den typiskt mindre eleganta kodstruktur som lätt kan uppstå vid exempeldriven regelkonstruktion: det är nästintill omöjligt att veta vad som är basfall, vad som är mest frekvent à priori (till skillnad från vad som är *grammatiskt möjligt*). Det typiska för stora program med ambitioner rörande hög täckningsgrad hos fri text verkar vara en sådan rörighet, i kontrast till grammatiker i en begränsad språkmodell. Med denna erkända brist på tydlighet är det inte desto mindre ett mycket fruktbart sätt att åstadkomma hög korrekthet.

Exempel på svårigheter i regelskrivningen

Somliga adverbialled kan som nämnts ha fler olika möjliga tolkningar beroende på kontext. Ibland kan endera tolkning vara associerad med en roll som valenskomplement, ibland finns sådan ledande valensinformation inte med. Ett adverbial som *i segment* kan ha mappningen *var* men om verbet är t.ex. sönderdela bör det snarare vara *hur*. Valensinformationen och tolkningen därav är tyvärr inte alltid perfekt.

I två grupper är ett typiskt fall som kan omfrågas med *hur* eller *var*. Om valenslexikonet vore idealiskt skulle hanteringen med fördel vara att ge motsvarigheten *var* som default och lita på att verbvalenshanteringen fungerar perfekt och lyckas finna precis de rätta undantagen. Men ibland är det lockande att redan för prepositionsfrasen i sig bestämma tolkning.

Generellt är det inte syftet att generera frågor som berör attribut i adverbialled utan hela led, vilket ofta innebär att just preposition och rektionshuvudord beaktas i PP-adverbial. Ytterligare ett tecken på att det är det praktiska som ändå fått styra är hur (framförställda) attribut ändå ges huvudrollen i många fall. Det gäller tydligt i fallet *med*-PP som annars måste omfrågas med det vagare *hur*, eller med pied piping. De *med*-adverbial som förekommer har en tendens att ha numeriska framförställda attribut *som med 5-0, med 123 procent* eller *med 90 deltagare*. I dessa fall skapas med några mönstermatchning några attributfokuserande frågor som *med hur mycket, med hur många procent* etc. *Med*-PP är en av flera grupper där det möjligen korrekta *hur* allt som oftast blir vagt och där pied piping-varianten *med vad* känns lämpligare.

Det är i regelkonstruktionen ofta värt att beakta vilken form ett substantiv som rektionshuvudord i en PP står i. Det visar sig att ett obestämt substantiv, speciellt när det ensamt fungerar som prepositionskomplement, är en indikator på att frågemotsvarigheten är annorlunda än normalt. *I vinkel* bör i enlighet med denna tumregel behandlas annorlunda än normalt: Frågemotsvarigheten blir närmast *hur* till skillnad från *i vinkeln* vilken tillåter tolkningen ”var”.

Om förändring av svarsformen under projektet

I de första versionerna fanns ofta en lösningsmodell som innebar att ett adverbial kunde ges flera olika *hv*-led som *hur/pied piping*. För många adverbial finns goda möjligheter till omfrågning på mer än ett sätt. Men i takt med att uppgiften började kräva en analys som involverade avgörande t ex om huruvida ett substantiv var konkret/abstrakt, motsvarar en händelse och annan inte alltid tillgänglig information blev de dubbla svaren för vissa adverbialslag snarare en

sorts garderingslösning. Speciellt prepositionerna behandlades en andra gång senare och det har mot slutet varit ett klart syfte att försöka ge enbart en lösning per frågeled. Detta har naturligtvis betydelse för en rättvisande utvärdering och eventuella jämförelser. Detta har betytt att PP-adverbial där det varit lockande att ange en sorts gardering *till en början → när/hur*, oftast har skrivits om till ett enda frågeled.

Det har verkat ofrånkomligt att med denna teknik samtidigt generera en hel del fel. Valet för uppgiften frågeledsbestämning blir att antingen att bara täcka in en liten del av alla förekommande adverbial perfekt, eller, vilket tillhör detta projektets grundförutsättningar, att göra försök för samtliga förekomster i text.

Oklara fall av frågeledsmotsvarigheter

En generell svårighet med frågeledsbestämning som har accentuerats i detta arbete är bestämmandet av vad som är det mest korrekta eller naturliga frågeledet. Denna typ av frågeställning har framträtt i de systematiska genomgångarna där huvudord mappats uttryckligen och när grupper (från FrameNet) med antaget liknande egenskaper har behandlats. Inte minst en gruppering inklusive satsadverbial och adverbfraser har medlemmar som verkar ha oklart beteende i sammanhanget.

Somliga rektionshuvudord som betecknar händelse med tidslig och rumslig placering kan vara svåra att sortera. *I en radiointervju* blir troligen med fördel *när*, men fungerar inte också *var*? *I* följd av ett obestämt substantiv i singular (ofta konkret) signalerar väldigt ofta ett specialfall: *i samtal, i dur, i slag* osv. Men här finns undantag som *i läsning* osv. där frågeledet kanske bättre motsvaras av pied piping, *i X* (där X som förekommer just så i programmet svarat mot en *hv*-motsvarighet av rektionen och kan vara både adverbiell och nominal, det avgörs i ett senare läge).

Men det finns också led som är fixerade uttryck och där den aktuella uppgiften, mappning till motsvarande frågeled, framstår som lite märklig, oavsett ledets placering. Detta är inte enbart något som gäller för satsadverbial utan för en hel del andra adverbialuttryck.

Till dags dato är ett exempel av många på ett sådant led som kan kännas igen som ett tidsadverbial och som kan omfrågas t.ex. med *när* eller liknande. Ändå är detta fasta uttryck en typ av information som inherent ter sig en aning bakgrundsartad (ännu mer specifikt skulle det utan övrig markering företrädesvis kunna räknas som *scen*-del av en *topik*). Vad som antyds här är att informationen i sig är mindre benägen att fungera som svar på realistiska frågor – även om den verkligen kan omfrågas. Idén skulle vara att förekomster av

sådana uttryck i text oftast utgör icke-rematiska led och inte normalt omfrågas. De skulle alltså kunna sorteras ut.

När det gäller det praktiska arbetet får behandling och försök till regelskrivning även för sådana begrepp som *till dags dato*, vilka är ganska frekventa, tidvis detta totala perspektiv att verka en aning ofruktbar och märklig. En möjlighet är alltså att över huvud taget inte beakta dessa fall, men som också klargjorts är det ändå möjligt att omfråga dem. Dessa gränsfall för omfrågbarhet tillhör de sidor av projektet som emellanåt utsätter det för en känsla av inexakthet.

Somliga substantiv av ’event’-typ, t.ex. *OS i Aten* svarar mot både *var* och *när*. Ett känt exempel från tidigare forskning är *’– När var det?’ ’I Moskva’*.
När PP-adverbial som inleds med t ex *från* studeras framgår ett fenomen som gäller för många PP-adverbial. De vanligaste förekommande *från*-adverbialen i text (SUC) bygger ett uttryck som är företrädesvis spatialt och kan omfrågas med *varifrån* eller en pied piping-variant: *från X* (där X bestäms genom komplementdelen).

Från Senneshyttan, Från TT, Från Tamanrasset, Från UD, Från baren

Från någon gång på 1870-talet är ett exempel på temporalt adverbial som troligen bäst omfrågas med *närifrån*.

Från alla epoker exemplifierar också en adverbialtyp som har en temporal betydelse, det vore rimligt att *närifrån* (eller *ifrån när*) används för detta fall. Intuitionen säger dock att *varifrån* – i en något mer abstrakt betydelse än tidigare – likaså kan användas även här. För att undvika de fel som kan uppstå vid det svåra urskiljandet av period- och händelse-substantiv från de som är spatiala är detta ett fall då en konsekvent mappning av samtliga *från*-adverbial (förutom fasta uttryck och satsadverbial) till en *hv*-frågetyp kanske är det lämpligaste.

Att samma adverbial har flera möjliga ’frågelösningar’ och att somliga frågeord har inte bara olika strukturtyper som svar – utan vad som kan vara olika semantiska kategorier – får vissa av *hv*-frågeleden att framstå som funktionellt ambiguösa. Rent allmänt innebär situationen annars naturligtvis att uppgiften är något enklare om flera lösningar godtas.

Tillbaka till uppgiftens karaktär

Den motsatta frågan till den som ställs i detta projekt, dvs. att istället reda ut svarsformen för en godtycklig fråga, har klart annorlunda förutsättningar än att välja frågeled för ett adverbialled. Det aktuella projektet antyder att svenska adverbialled generellt har ett fåtal och ofta ett enda motsvarande *hv*-frågeled även om det alltså kan vara svårt att alltid avgöra vad det är.

Om det på motsvarande sätt hade varit så att varje *hv*-fråga endast kunde ha en enda bestämd form så skulle det förmodligen vara möjligt att mekaniskt dra långtgående slutsatser angående texters svarspotential. Emellertid följer de svar som vanliga *hv*-frågor kan ha i t.ex. FAQ-uppställningar olika former och är ofta inte (elegant) komprimerbara till ett enda satsled utan består t.ex. av flera satser. Generellt antas varje *hv*-fråga kunna besvaras genom många olika strukturer. Medan ett enskilt adverbialled svarar mot ett fåtal *hv*-ord så kan varje *hv*-frågas svar oftast anta många olika former, trots likartat informationsinnehåll.

Om idén fullföljs, att samma (svars-)information kan uttryckas komprimerat eller över många satser på ett sätt som kan fångas regelmässigt, så väcks ändå förhoppningen om att mekaniskt kunna dra slutsatser om ifall en frågas svars eventuella befintlighet i en text, utgående från den form det borde ha. Här finns då tanken att transformationer av en informationsportion (ett frågesvar) tillsammans med lexikala utbyten ska kunna producera alla möjliga utseenden för denna specifika information. Detta är en hållning som i så fall försöker stävja och fånga in den *frihet i språkligt uttryckssätt* som varit föremål för många språkvetenskapliga uppsatser. Vore det möjligt att givet en fråga och ett visst svar på denna fråga generera samtliga de uttrycksformer svaret kan anta i svensk textform?

En förbipasserad fråga gäller om alla olika *hv*-frågor kan ha ett svar som kan besvaras med ett enda adverbialled, eller om den genomgång från adverbialperspektivet som skett här har missat någon möjlig *hv*-fråga helt. Ett antagande här är att medan de olika adverbialleden inte enligt den aktuella mappningen har kopplats till samtliga tänkbara *hv*-ord så antas samtliga sådana frågor kunna ha svar som uttrycks, eventuellt pronominaliserat i ett enda satsled.

Citerade arbeten

Abney, S. (1991). Parsing by Chunks. In *Principle-Based Parsing.* Kluwer Academic Publishers.

Borin, L., Dannélls, D., Forsberg, M., Toporowska Gronostaj, M., & Kokkinakis, D. (2010). The Past Meets the Present in the Swedish FrameNet++. *14th EURALEX International Congress*, (ss. 269-281).

Borin, L., Forsberg, M., & Lönngren, L. (2008). *SALDO 1.0 (Svenskt associationslexikon version 2).* Göteborg: Språkbanken, Göteborgs universitet.

Diderichsen, P. (1946). *Elementær Dansk Grammatik.* Köpenhamn: Gyldendahl.

Ejerhed, E., Källgren, G., & Brodda, B. (2006). *Stockholm-Umeå corpus version 2.0.* Institutionen för Lingvistik, Stockholms universitet, Institutionen för Lingvistik, Umeå universitet.

Forsbom, E. (2008). Good Tag Hunting: Tagability of Granska Tags. i J. Nivre, M. Dahllöf, & B. Megyesi, *Resourceful Language Technology: Festschrift in Honor of Anna Sågvall Hein, ACTA UNIVERSITATIS UPSALIENSIS Studia Linguistica Upsaliensia 7.* Uppsala.

Hellberg, Staffan;. (2003). Varför inte prepositionsobjekt? i C. F. Lars-Olof Delsing, *Grammatik i fokus. Festskrift till Christer Platzack* (ss. 47–53). Lund: Lunds universitet.

Hultman, T. G. (2003). *Svenska akademiens språklära.* Stockholm: Svenska akademien. Norstedts ordbok distributör.

Lexin – Svenska ord. (1998). Norstedts Akademiska Förlag.

Ljung, M., & Ohlander, S. (1971). *Allmän grammatik.* Malmö: Gleerups.

Nationalencyklopedins ordbok. (1995–96). Höganäs: Bra Böcker.

Nationalencyklopedins ordbok. (1995–96). Höganäs: Bra Böcker.

Øvrelid, L. (2008). *Argument Differentiation. Soft constraints and data-driven models (Doktorsavhandling).* Göteborg: Göteborgs universitet.

Ross, J. R. (1967). *Constraints on variables in syntax.* Doktorsavhandling, MIT.

Teleman, U. (1974). *Manual för grammatisk beskrivning av talad och skriven svenska.* Lund: Studentlitteratur.

Wilhelmsson, K. (2011). Automatic Question Generation from Swedish Documents as a Tool for Information Extraction. *Proceedings of the 18th Nordic Conference of Computational Linguistics.* Riga.

Wilhelmsson, K. (2008). Heuristic Schema Parsing of Swedish Text. *Proceedings of SLTC 2008.* Stockholm.

Wilhelmsson, K. (2010). *Heuristisk analys med Diderichsens satsschema - Tillämpningar för svensk text (doktorsavhandling).* Göteborgs universitet: Institutionen för filosofi, lingvistik och vetenskapsteori.

Appendix

Adverb **AB** *inte*	Particip **PC** *utsänd*
Determinerare **DT** *denna*	Partikel **PL** *ut*
Frågande/relativt adverb **HA** *när*	Egennamn **PM** *Mats*
Frågande/relativ determinerare **HD** *vilken*	Pronomen **PN** *hon*
Frågande/relativt pronomen **HP** *som*	Preposition **PP** *av*
Frågande/relativt possessivt pronomen **HS** *vars*	Possessivt pronomen **PS** *hennes*
Infinitivmärke **IE** *att*	Grundtal **RG** *tre*
Interjektion **IN** *ja*	Ordningstal **RO** *tredje*
Adjektiv **JJ** *glad*	Subjunktion **SN** *att*
Konjunktion **KN** *och*	Utländskt ord **UO** *the*
Substantiv **NN** *pudding*	Verb **VB** *kasta*

Tabell 8 Varje löpord i SUC 2.0 är uppmärkt med en av ovanstående ordklasstaggar, oftast kombinerat med särdragsvärden från Tabell 9[23]. *Hämtat från Avh.*

Särdrag	*Möjliga särdragsvärden*	*Ordklasser där särdraget är tillämpbart*
Genus	**UTR** Utrum **NEU** Neutrum **MAS** Maskulinum	DT, HD, HP, JJ, NN, PC, PN, PS, (RG, RO)
Numerus	**SIN** Singularis **PLU** Plural	DT, HD, HP, JJ, NN, PC, PN, PS, (RG, RO)
Bestämdhet	**IND** Obestämd **DEF** Bestämd	DT, (HD, HP, HS), JJ, NN, PC, PN, (PS, RG, RO)
Kasus	**NOM** Nominativ **GEN** Genitiv	JJ, NN, PC, PM, (RG, RO)
Verbform	**PRS** Presens **PRT** Preteritum **SUP** Supinum **INF** Infinitiv	VB
Diates	**AKT** Aktiv **SFO** S-form (passiv eller deponensform)	
Modus	**KON** Konjunktiv	
Participform	**PRS** Presens **PRF** Perfekt	PC
Kompareringsform	**POS** Positiv **KOM** Komparativ **SUV** Superlativ	(AB), JJ
Pronomenform	**SUB** Subjektsform **OBJ** Objektsform	PN
Sammansättningsform	**SMS** Sammansättningsform	Nästan alla ordklasser (i teorin)

Tabell 9 Möjliga särdrag och särdragsvärden i möjliga kombinationer efter *Manual of the Stockholm Umeå Corpus version 2.0* (Ejerhed, Källgren, & Brodda, 2006). Koder inom

[23] Härutöver kommer markering för interpunktioner (*MAD* för textmeningsavgränsare som ”.” och ”!” eller *MID* för andra som ”,” och ”;”) samt parvisa avgränsare (citattecken, parenteser): *PAD*.

parentes innebär att särdragen är tillämpliga på bara en del av ordklassens medlemmar eller att bara en del av särdragsvärdena är möjliga. *Hämtat från Avh.*

Något om lexikala resurser för substantivklassificering

SweFN (Borin, Dannélls, Forsberg, Toporowska Gronostaj, & Kokkinakis, 2010) är en rik digital lexikonstruktur bland Språkbankens fritt tillgängliga resurser som här har undersökts med syftet att klassificera främst substantiv. I samlingen förekommer ca 460 semantiska kategorier som har minst ett substantiv bland kategoriexemplen. Dessa sorterades först på kategorinamn med tillhörande exempelord (en kontroll mot SALDO användes för att utesluta kategorier utan möjliga substantivexempel).

Court_examination: korsförhör, korsförhörande, utfrågning, utfrågande

Den fråga som nu framträdde var framför allt: *hur enhetligt fungerar varje semantisk kategori i SweFN i fråga om gemensam omfrågbarhet?* För att undersöka detta gicks de drygt 470 kategorierna med exempelord igenom i ett kategoriseringsgränssnitt som byggdes. I denna sorteringsprocess visades exempelorden tillsammans med en prototypisk PP-konstruktion ”I min/mitt…” följt av de aktuella exempelorden. Varje kategori sorterades därmed i en omfrågningskategori, i detta fall skapades ca 18 sådana omfrågningkategorier.

En viktig hypotes på förhand var att distinktionen konkreta/abstrakta substantiv skulle ha stor betydelse för hur ett PP-format adverbial med svag preposition och detta rektionshuvudord skulle fungera. En enkel hypotes var att med en preposition som *i* skulle prepositionsfraser med konkret rektionshuvudord ges den för *i*:s prototypiska rumstolkningen och efterfrågas med hjälp av *var*. Resultatet visade en mer komplex bild än så. Somliga av de semantiska kategorierna med många exempel av typen konkret substantiv antogs emellertid verkligen fungera så:

Accoutrements:	*accessoarer, armband, armbandsur, berlock, bijouterier…*
Optical_image:	*skugga, skuggbild, silhuett, silhuettering, spegelbild*
Part_edge:	*kant, utkant, ytterkant, periferi, bryn, skogsbryn*

I flera semantiska kategorier förekommer substantivexempel med abstrakta substantiv som semantiskt sett betecknar exempelvis händelser, vilka rimligen har både tids- och rumsaspekter, som *var* och *när*, ibland även t.ex. *hur*:

Vehicle_landing:	*landa, landning, mellanlandning, månlandning, nödlandning*
Using:	*tillämpande, tillämpning*
Undergoing:	*undergående, underkastande, genomgående, genomlidande*

Traversing:	*bestigning, uppstigande, uppstigning, cirkulering…*
Transfer:	*överförande, överföring*
Touring:	*sightseeing, rundtur, sevärdhet, turist, turism*
Telling:	*underrätta, försäkran, informering, kommuniké, meddelande*

Många semantiska kategorier, bl.a. sådana med participsubstantiv hamnade i en gruppering med oklar omfrågningsbarhet. Dessa kan antas vara sådana som rimligen omfrågas direkt.

Willingness: benägenhet

I ett fall som det tänkbara adverbialet ’i min benägenhet’ får motsvarande omfrågning, om över huvud taget någon, falla tillbaka på det prototypiska *var* eller *i vad* (eller *vad + strandad preposition*). Denna gruppering har tillsammans med flera andra märkts som ’icke benägen att omfrågas’, vilket kanske är den bästa beskrivningen för praktiska syften.

Efter kategoriseringen hade vidden av svårigheten i uppgiften undersökts. De substantiv var samtliga sådana som förekommer i SALDO. Det innebär att den kategorisering som gjorts kommer att bli resultatet för en tänkt allmän svag preposition som *i*.

Som språkteknologisk resurs har SweFN visat flera andra kvaliteter då den exempelvis innehåller substantivgrupperingar som svarar mot mänskliga referenter – en användbar information för språkteknisk subjektsbestämning, frågeledsbestämning m.m.

Undersökningen av de semantiska kategorierna visade även att många hade medlemmar som troligen fungerar ganska olika och där den semantiska kategorin i vissa fall endast spelar rollen som en allmän association till medlemmarna. I exemplet *touring* nedan får antas att ’på turisten’ omfrågas med *var* medan ’på rundturen’ kan vara t.ex. *var* eller *när*.

Touring: sightseeing, rundtur, sevärdhet, turist, turism…

SweFN har uppenbart inte skapats för att just underlätta frågegenerering. Det är inte desto mindre mycket bra att somliga kategorier har kunnat hjälpa till att lämna viktiga basfall till kategorier som personer, platser, tidpunkter och konkreta ting. Användningen av SweFN i programmet just nu är sparsam men den förekommer för att avgöra konkreta substantiv.

GU-ISS, Forskningsrapporter från Institutionen för svenska språket, är en oregelbundet utkommande serie, som i enkel form möjliggör spridning av institutionens skriftliga produktion. Det främsta syftet med serien är att fungera som en kanal för preliminära texter som kan bearbetas vidare för en slutgiltig publicering. Varje enskild författare ansvarar för sitt bidrag.

GU-ISS, Research reports from the Department of Swedish, is an irregular report series intended as a rapid preliminary publication forum for research results which may later be published in fuller form elsewhere. The sole responsibility for the content and form of each text rests with its author.

Forskningsrapporter från institutionen för svenska språket, Göteborgs universitet
Research Reports from the Department of Swedish

ISSN 1401-5919

www.svenska.gu.se/publikationer/GU-ISS