Heuristisk analys med Diderichsens satsschema

GOTHENBURG MONOGRAPHS IN LINGUISTICS 40

Heuristisk analys med Diderichsens satsschema

Tillämpningar för svensk text

Kenneth Wilhelmsson

Nationella forskarskolan i språkteknologi, GSLT

GÖTEBORGS UNIVERSITET

Avhandling för filosofie doktorsexamen i allmän språkvetenskap
Göteborgs universitet 2010
Disputationsupplaga för tryck i A5-format

Tryckt av Reprocentralen, Humanistiska fakulteten, Göteborgs universitet

Distribution:
Institutionen för filosofi, lingvistik och vetenskapsteori, Göteborgs universitet
Box 200, 405 30 Göteborg

Ph.D. Thesis in linguistics at University of Gothenburg 2010

Title: Heuristisk analys med Diderichsens satsschema – Tillämpningar för svensk text
English title: Heuristic Analysis with Diderichsen's Sentence Schema – Applications for Swedish Text
Author: Kenneth Wilhelmsson
Language: Swedish (including summary in English)
Department: Department of Philosophy, Linguistics and Theory of Sciences

Abstract

A heuristic method for parsing Swedish text, *heuristic schema parsing*, is described and implemented. Focusing on main clause (*primary*) analysis, a collection of licensing techniques for removing non-primary verb candidates is employed, leaving e.g. the primary verbs, particles and conjunctions (bounded key constituents) that delimit the content of the fields in Diderichsen's sentence schema. Hereby, the subsequent identification of constituents which do not have an upper bound on their length (subject, object/predicatives and adverbials) can be identified relying to a lesser on extent explicit pattern matching, and more on different heuristic rules. For phrase type identification and delimitation of these constituents, when adjacent to each other, a novel chunking technique, *rank-based chunking*, is applied. Following this, a series of further rules merge chunks into larger ones, aiming at a final number of nominal chunks compatible with the valency information of the main verb. The aim is to identify *full* nominal and adverbial constituents, including post-modifiers. The implementation uses the *Stockholm Umeå Corpus 2.0*, a corpus which is balanced for different genres in published Swedish text. SUC*'s* tagset is also used unmodified in part-of-speech tagging which enables the program to deal with input text. The functional parsing, which includes no explicit language-defining grammar component is carried out technically using an object-based representation of clause structure.

Although output formats and types of evaluations of correctness are very different in parsers for Swedish text, it is claimed that the manual approach presented can provide high accuracy, which can be improved given more time for development.

The thesis work also includes two prototype applications, both requiring high accuracy of the sort of functional syntactic analysis described here. The first one is an implementation of automatic syntactic fronting in the area of text editing for Swedish, where the user is presented with a syntactically analyzed copy of her writing, from which paraphrases easily can be generated. The second application is in the field of natural language query systems and produces questions with answers from an arbitrary declarative input text. This prototype incorporates a text database from Swedish *Wikipedia*, and investigates primarily generation of *WH*-questions formed via fronting of unbounded primary constituents. The questions are generated as a text is opened and thus permits users to only ask the available ones, thus aiming at a high precision value.

Keywords: Diderichsen's sentence schema, positional grammar, field grammar, licensing techniques, *Stockholm Umeå Corpus*, schema parsing, rank-based chunking, syntactic fronting, paraphrasing, question generation, natural language query systems, Swedish *WordNet*

See also Summary in English *at the end of this thesis*

Doktorsavhandling i allmän språkvetenskap, Göteborgs universitet 2010

Titel: Heuristisk analys med Diderichsens satsschema – Tillämpningar för svensk text
Engelsk titel: Heuristic Analysis with Diderichsen's Sentence Schema – Applications for Swedish Text
Författare: Kenneth Wilhelmsson
Språk: Svenska (med sammanfattning på engelska)
Institution: Institutionen för filosofi, lingvistik och vetenskapsteori

Sammandrag

En heuristisk metod för parsning av svensk text, *heuristisk schemaparsning,* med implementation beskrivs. Med fokus på huvudsatsanalys används en samling licensieringstekniker för att utesluta icke-primära kandidater till de längdbegränsade (eng: *bounded*) nyckelkomponenter som avgränsar fält och andra utrymmen i Diderichsens satsschema. Härigenom kan de funktionella konstituenter som är (potentiellt) obegränsade i längd (eng: *unbounded*), subjekt, objekt/predikativ och adverbial, identifieras genom att i lägre grad använda explicit matchning av flerordsled och istället tillämpa olika heuristiska regler. För frastypsbestämning och avgränsning av dessa konstituenter, när de är angränsande, presenteras först en ny segmenteringsmetod, *rangbaserad chunkning*. Denna segmentering följs av en serie möjliga sammanfogningar som syftar till att nå ett antal nominala led som är kompatibelt med valensen hos satsens huvudverb. Målsättningen för denna metod är identifikation av *hela* nominala och adverbiella led, inklusive efterställda attribut. Detta avhandlingsprojekt baseras på *Stockholm Umeå Corpus 2.0* som speglar olika genrer av svensk publicerad text. Dess tagguppsättning används också omodifierad i en ordklasstaggare som möjliggör hantering av valfri textinput. Den inre representationen av en textmening, under denna funktionella syntaxanalys som inte innehåller någon explicit språkdefinierande grammatikkomponent, är objektbaserad.

Även om utdataformat och förutsättningar för korrekthetsutvärderingar varierar mycket för svenska parsningsprojekt, hävdas att ansatsen kan ge hög korrekthet, vilken kan förbättras om mer tid ägnas åt det manuella regelskrivandet.

Avhandlingsarbetet inkluderar även två prototyptillämpningar som båda kräver hög korrekthet av den analysform som här produceras. Den första är en implementation i området ordbehandling där en användare ges möjlighet att automatiskt parafrasera skrivna textmeningar då syntaxanalys av dessa visas. Den andra applikationen som presenteras tillhör området natural language query systems och genererar automatiskt frågor till en godtycklig införd text. Denna prototyp inbegriper textdatabasen från svenska *Wikipedia* och undersöker främst generering av *hv*-frågor som bildas genom spetsställning och mappning till frågeord. Frågegenereringen sker när en text öppnas och tillåter frågor från användaren med speciellt fokus på precionvärdet – hög korrekthet på svaren givet frågorna.

Nyckelord: Diderichsens nordiska satsschema, positionsgrammatik, fältgrammatik, licensieringstekniker, *Stockholm Umeå Corpus*, schemaparsning, rangbaserad chunkning, spetsställning, parafrasgenerering, frågegenerering, naturligt språk-frågesystem, svenska *WordNet*

Tack

Jag har under min tid som doktorand befunnit mig i en privilegierad situation. Jag vill här försöka nämna många av de som bidragit till detta avhandlingsarbete. Först och främst vill jag tacka mina handledare. Min huvudhandledare Robin Cooper har varit till stöd under hela arbetets gång, diskuterat det mesta och tipsat om litteraturreferenser och lingvistisk teori och metod i de delområden som arbetet berör. Min bihandledare Dimitrios Kokkinakis har med sin erfarenhet generöst delat med sig, både ifråga om teoretiska och praktiska tips och väsentligheter typiska för uppgifterna i automatisk syntaxanalys och forskningsfältet informationssökning.

I det praktiska arbetet har jag fått hjälp och goda råd från många medarbetare. Angående den slutversion av ordklasstaggare som förekommer (Kapitel 4) har jag kunnat fråga Leif Grönqvist, som med mycket stor hjälpsamhet ställt upp och svarat på frågor. Jag vill också speciellt tacka Viggo Kann som hjälpte till att ordna en kurs för mina ändamål, Ola Knutsson och Jonas Sjöbergh på KTH för mycket tillmötesgående och hjälpsamhet vid olika praktiska delar – ordklasstaggning och chunkning. När det gäller syntaktisk valensdata som här används är jag också mycket nöjd med att ha fått ta del av Institutionen för svenska språkets resurser – det har framför allt här rört sig om databasförlagor till ordböcker. Tack särskilt till Maria Toporowska Gronostaj, Lars Borin och till Sven-Göran Malmgren för ett trevligt bemötande och bra svar på mina frågor gällande valensdata för svenska och rättighetsfrågor. Från Institutionen för svenska språket har jag även kunnat konsultera Maia Andréasson för mina frågor om adverbial som ju också är relevant i detta arbete (Kapitel 2 och 3). I fråga om subjektsidentifikation på grundval av bl.a. animathet har jag fått svar på frågor av Lilja Øvrelid som ägnat sig åt detta område. Hans Landqvist var hjälpsam i slutskedet med en diskussion om *Mamba*-relaterade ställningstaganden.

Från Stockholms universitet har man hjälpt mig att nå Janne Lindberg vars magisteruppsats i datorlingvistik (skriven tillsammans med Carin Svensson) har varit en unik föregångare ifråga om automatisk omformulering av svensk text genom spetsställning (Kapitel 5). Eftersom detta arbete i mycket hög grad har varit beroende av *Stockholm Umeå Corpus* har jag haft anledning att kontakta de idag ansvariga för denna resurs. Jag vill tacka Sofia Gustafsson Capková, Eva Forsbom och Britt Hartmann för att ha tagit sig tid när jag kontaktat dem av olika korpusrelaterade anledningar. Jag har fått hjälp av Institutionen för lingvistik i Stockholm flera gånger. Därifrån har man skaffat fram och skickat en särskild rapport av Gunnel Källgren åt mig.

Jag har fått relevanta och intressanta kommentarer från skilda håll i fråga om ansatsen i detta arbete – med risk för att glömma någon vill jag i alla fall nämna Lars Ahrenberg, Lars Borin, Torbjörn Lager, Jens Allwood, Magnus Gunnarsson, Joakim Nivre, Svetoslav Marinov, Elisabet Engdahl, Benny Brodda, Sofie Johansson Kokkinakis, Maria Toporowska Gronostaj, Leif Grönqvist, Östen Dahl, Harald Hammarström, Karin Cavallin, Ulrika Kvist Darnell, Staffan Larsson och Sören Sjöström. Jag vill tacka Åke Viberg för möjligheten att undersöka den svenska versionen av *WordNet* (Kapitel 5). I detta sammanhang vill jag också gärna lyfta fram den hjälpsamhet som Johan Dahl och Robert Andersson visade i det praktiska extraktionsarbetet.

För hjälp med stora och små datorrelaterade frågor under perioden vill jag tacka Per Olofsson, Robert Andersson och Peter Nilsson. När det gäller administrativa frågor inför disputationen har jag kunnat fråga Åsa Abelin, Hans Vappula, Pia Gårdmo och Ann Mari Teiffel som varit hjälpsamma. Jag vill också tacka personalen på Reprocentralen: Christina Gudmundsson och Henrik Rundqvist, Humanisten, för hjälpsamhet och svar på många frågor inför tryckningen.

Institutionen för lingvistik, som under större delen av perioden har varit namnet på min hemvist, vill jag tacka för den forskarmiljö, med hjälpsamma kollegor som erbjudits mig, men även de nya kollegorna från den nya sammanslagna institutionen som har visat intresse.

Inför slutredigeringen av föreliggande text har jag haft hjälp genom kommentarer från genomläsningar av Pierre Gander, Sofie Johansson Kokkinakis och Maria Toporowska Gronostaj.

Om avhandlingstexten är läsbar och har få skriv- och språkfel så beror det på att texten i slutskedet har språkgranskats av Ylva Byrman. Eventuella kvarvarande fel är författaren själv ansvarig för.

På den mer personliga sidan går mitt tack till far, mor, bror, släkten och mina vänner.

Slutligen vill jag tacka Nationella forskarskolan i språkteknologi (GSLT) för finansieringen som dessutom lägligt förstärktes genom stipendier från Helge Ax:son Johnsons stiftelse samt från Adlerbertska stiftelserna.

Avhandlingens vokabulär och konventioner

Det nedanstående är en uppställning av några termer, inklusive förkortningar, som används i texten. De flesta av dessa tillhör vedertagen terminologi inom språkvetenskapen. Se även *Index* längst bak.

Adverbiell: Ett adverbiellt led uppbär adverbialroll, motsats till nominal, och är typiskt en adverb- eller prepositionsfras.
Begränsad: Av begränsad (potentiell) längd. Termen används i detta avhandlingsprojekt i analogi med engelska *bounded*. I satsschemaanalysen är de positionsinnehåll som bara kan rymma ett eller ett fåtal ord begränsade (icke-rekursiva). Häribland finns verb, partiklar, satssamordnande konjunktioner m.fl. (Detta har *inte* med *bundet adverbial* att göra.)
Efterdel: I detta arbete: beteckning för hela fältet efter finitet i en icke-hjälpverbskonstruktion. Detta fält hanteras utan att eventuell gräns mellan mitt- och slutfält görs.
Finit: Finit verb, ett sådant som normalt krävs för att skapa en fullständig sats och står i presens, preteritum eller imperativ.
Fristående led: Att ett led är fristående innebär i detta arbete att det innehar en egen position i satsschemat, eller skulle kunna ha det (om t.ex. två objekt delar på N-positionen). Det är alltså inte en del av något annat satsled på den aktuella nivån.
Fundament: Den första positionen (före det finita verbet) i det normala satsschemat, dit ett led kan spetsställas (fundamenteras) från sin kanoniska position, se även *Satsbas*.
Förfält: En position i det utökade satsschemat som föregår fundamentet och typiskt innehåller en konjunktion men även längre initiala annex: *Till Paris, de ville dit nu.*
***HV*-frågor:** *Frågeordsfrågor*. Frågor som inleds med ett *hv*-ord kan motsvara ett satsled. Av engelska *wh-questions*: *Vad, var, varför, vem, när* m.fl. Jfr V1-frågor.
Infinit: Diderichsens danska term, icke-finit verb. Förekomst kan indikera hjälpverbskonstruktion och oftast avgränsat mittfält.
Licensiering: (främst av verb) Tillåtande av t.ex. verbförekomster. Genom licensiering av verb på underordnade nivåer klargörs de återstående (primära) verben. Därmed får licensiering i den aktuella metoden nästan betydelsen *borttagning* (av kandidat). Se vidare i *2.3 Primärfinitsidentifikation*.
***Mamban*:** Manualen för syntaktisk annotering som kallas *Mamban* är egentligen två olika skrifter från lundaprojekten om tal- och skrivsyntax, den äldre, även kallad *Manualen*, (Loman och Jörgensen 1971) och den nyare (Teleman 1974). Den nyare som fokuserar mer på syntaxanalys åsyftas här om inget annat nämns.

***NEO*:** En vedertagen förkortning för *Nationalencyklopedins ordbok* (1995–96). I detta avhandlingsprojekt används bitvis valensinformation ur en databas som ligger till grund för denna ordbok, och till *Svensk ordbok* (2009).
Nominal: Nominala led är de strukturer som uppbär roller som subjekt och objekt. Enligt en tolkning (av Diderichsen) kan även predikativ ingå (se Formel 4).
Obegränsad: Som obegränsade (*unbounded*) satsled räknas de nominala och adverbiella leden: subjekt, objekt/predikativ och adverbial. Dessa har (potentiellt) obegränsad längd.
Precision: *Precision*-värdet är ett mått på hur felfri en mängd uppmärkningar är, hur stor del av den som innehåller enheter som verkligen borde finnas däri (jfr: *recall*).

$$Precision = \frac{Antal\ korrekta\ uppmärkningar}{Totalt\ antal\ uppmärkningar\ som\ systemet\ ger\ som\ resultat}$$

Formel 1 *Precision*-värdet kan beskrivas som hur 'ren' resultatmängden är om varje felaktigt svar ses som förorenande av denna.

Primär: På huvudsatsnivå i satslösning. I *Den som vi valde vann* är *vann* primärt finit och hela segmentet *den som vi valde* är primärt subjekt.
Recall: *Recall*-värdet är ett mått på hur många av de enheter som borde finnas i en svarsmängd som verkligen finns där, och hur många som missats att välja ut.

$$Recall = \frac{Antal\ korrekta\ uppmärkningar\ som\ gjorts}{Totalt\ antal\ korrekta\ uppmärkningar\ som\ borde\ markeras}$$

Formel 2 *Recall*-värdet visar hur stor del av de egentliga svaren som systemet har i sin mängd av uppmärkningar.

Rektion: Komplement, här speciellt prepositionskomplement, t.ex.: *vid dörren, mot att rösta*.
Rekursiva led: Obegränsade (*unbounded*) led, de led som funktionellt uppbär subjekt, objekt/predikativ och adverbial. Strukturmässigt är det frastyper som NP, PP m.fl.
Satsbas: Fundamentinnehåll (används t.ex. i SAG).
Satsled: I texten betyder *led* förekomst av funktionell kategori och är här begränsade eller obegränsade i längd. Används av SAG ungefär som *satsdel*.
***SAG*:** *Svenska Akademiens grammatik* (Teleman, Hellberg och Andersson 1999)
S-enhet: Textenhet i *Stockholm Umeå Corpus* som svarar dels mot textmeningar bestående av en eller flera huvudsatser, men därutöver också mot rubriker och andra fristående led som inte nödvändigtvis avgränsas av stor bokstav och stort skiljetecken (*Som om det räckte, Ny lag*). Dessa enheter är grundläggande i detta arbete och analyseras en åt gången.

Spetsställning: Spetsställning innebär att ett led flyttas från sin ursprungsposition, enligt satsschemat till fundamentet (Adverbialspetsställning: Och *vi ska inte gå → Och inte ska vi gå)* och kallas även *fundamentering*. Termen *topikalisering* används dock inte enbart med denna betydelse.
SUC*: *Stockholm Umeå Corpus, här används genomgående version 2.0 (Ejerhed, Källgren och Brodda 2006).
V1-frågor: Fundamentlösa frågor (ja/nej-frågor): *Tänker ni odla vete här?*

Tecknet ←//→ står här för icke-ekvivalens i betydelse, t.ex. *Han sprang nog* ←//→ *Nog sprang han***.**

Se även *Index* i slutet av avhandlingen.

Innehållsförteckning

1 Inledning

När tillvägagångssättet för funktionell grammatisk analys av svenska beskrivs i välformulerade läroböcker som i Tabell 1 är det i form av en stegvis punktlista. Varje textmening, som t.ex. *Eftersom det regnade hade de tagit paraplyet i morse* kan undersökas med en algoritm som ser relativt likartad ut i litteraturen.

1 *När man gör satslösning i primära satsdelar är det lämpligt att börja med att leta rätt på huvudsatsens finita verb*. s. 47 i Josefsson (2001)	*Eftersom det regnade **hade** de tagit paraplyet i morse*
2 Om det är fråga om en hjälpverbskonstruktion som i exemplet identifieras även resten av verbkedjan, dvs. huvudsatsens huvudverb.	*Eftersom det regnade **hade** de **tagit** paraplyet i morse*
3 *Har man noterat verbfrasens olika delar kan man i regel finna subjektet genom den s.k. subjektsfrågan "vem/vad är det som + verbfrasen?"* s. 28 (Stroh-Wollin 1998)	*Eftersom det regnade **hade de tagit** paraplyet i morse*
4 *Det direkta objektet svarar på frågan Vad/vem + predikat + subjekt?* s. 47 (Josefsson 2001)	*Eftersom det regnade **hade de tagit paraplyet** i morse*
5 Sent i analysen eftersöks adverbial, varav TSR-adverbial besvarar *frågor som när, var, hur och varför.* s. 57 (Stroh-Wollin 1998)	***Eftersom det regnade hade de tagit paraplyet i morse***

Tabell 1 Den primära satslösningen sker i läroböcker genom en frågelista rörande satsens betydelse.[1]

Om de ovanstående punkterna får utgöra riktlinjer blir det tydligt att det är en metod som för alla steg kräver en viss insikt: den som gör övningen måste ha *förstått* satsen för att göra analysen. Det gäller först att veta vad som är huvudsatsens finita verb, i punkt 1, men också att kunna besvara frågor som verkar röra själva betydelsen. Betyder det att det är omöjligt att göra en liknande analys som ett förstasteg i en datoriserad parsning, som inte har denna förståelseförmå-

[1] Det bör klargöras att de nämnda läromedlen också ger viss positionsmässig ledning för analysen, exempelvis att subjektet oftast finns i anslutning till finitet, även om det inte är placeringen som definierar de olika satsleden.

ga? Denna avhandling som handlar om funktionell grammatisk parsning ställer denna fråga samt frågan om en sådan metod rentav kan vara fördelaktig.

Den metod som nordistiken erbjuder för satslösning på rent form- och placeringsmässiga grunder kommer från den danske professorn Paul Diderichsen, vars satsschema innebar ett viktigt bidrag till den s.k. traditionella grammatikbeskrivningen av nordiska språk. Eftersom analysen som här beskrivs är maskinell kan den alltså inte begagna sig av t.ex. 'subjektsfrågan' (*vem/vad + predikat?, dvs.* Vem/vad *hade tagit paraplyet?)* enligt Tabell 1 för att bestämma satsers subjekt. Till skillnad från flertalet implementationer för svensk syntaxanalys är den ändå mer lik nämnda skolövning eftersom den utgår från *primära finita verb* (dvs. finita verb på huvudsatsnivå) i huvudsatser och i finita verbfraser som är samordnade på samma nivå. Satsschemat kommer in i bilden som den självklara orienteringskarta en mekanisk analys behöver för en analys av svenska utgående från ordnings- och formkriterier. Med identifikation av primära finita verb som förstasteg möjliggörs en metodik där exakt matchning av andra led på huvudsatsnivå inte alltid är nödvändig. För att avgöra vilken konstituent som t.ex. finns mellan finit hjälpverb och infinit huvudverb, dvs. i *mittfältet* (*nexusfelt*, se Tabell 2), blir det här ofta möjligt att säga det genom uteslutning. Metoden startar således med analys på huvudsatsnivå – att nästan samma metodik kan användas för lägre nivåer är en viktig hypotes. Se schemat för bisatsnivå, Tabell 3. Det ska sägas att satsschemamodellen för nordiska språk förmodligen är omtyckt mycket p.g.a. sin enkelhet och kompakthet och det vore önskvärt om projektet inte komplicerar denna pedagogiska sida.

När det gäller grammatisk terminologi är förhoppningen att detta arbete i möjligaste mån ska begagna vedertagna begrepp som *fundament* på ett så okomplicerat och otvetydigt sätt som möjligt, och oftast i överensstämmelse med framför allt *Svenska Akademiens grammatik* (Teleman, Hellberg och Andersson 1999), hädanefter kallad SAG, samt nyare läromedel. Det finns dock några smärre skillnader, vilka kommer att nämnas. Som huvudsaklig målsättning i arbetet med själva analyskomponenten finns en så hög korrekt täckningsgrad av svensk text som möjligt – med hänsyn taget till hur textmeningar är fördelade över syntaktiska former. Detta innebär att ovanliga men korrekta syntaktiska variationer generellt har fått lägre prioritet här, även om korrekt analys också av dessa måste vara del av en slutlig målsättning.

1.1 Diderichsens satsscheman

Diderichsen presenterade sin beskrivning av den danska satsens topologi vid det åttonde nordiska filologimötet i Köpenhamn, 12 augusti 1935. En resumé av detta föredrag blev tryckt året därpå med titeln *Prolegomena till en metodisk*

dansk Syntax och återfinns i *Helhed og Struktur* (Diderichsen 1966). Från de rent textmässiga beskrivningarna av satsledens topologiska ordning som fanns där och senare i hans avhandling om satsbyggnaden i Skånelagen (Diderichsen 1941) har Henriksen (1986) skisserat de förlagor till satsschema som dessa motsvarar. Henriksen (1986) ger denna och annan upplysande information om åren kring satsschemats uppkomst. Det verkar som den första tryckta versionen av själva satsschemat, *tabellgrafiskt*, kom först ungefär tio år senare i *Elementær Dansk Grammatik* (Diderichsen 1946), s. 186. I boken finns schemat med som en pedagogisk översikt över ordningen hos satsleden.

Idag är huvudsatsschemat vedertaget för beskrivning av de nordiska språken. För svenska förekommer det nu i de flesta moderna 'traditionella' grammatikböcker på grundnivå. Speciellt framskjuten plats har satsschemat hos Holm och Larsson (1980). Schemat har oftast omkring sju positioner, fördelade på tre huvudsakliga fält – fundament, mittfält (nexusfält), och slutfält (innehållsfält), ibland med tilläggspositioner för en satsinledande konjunktion före fundamentet som *og* nedan (*Och* vi gick hem, dvs. forbinderfelt/förfält) eller för verbpartiklar (som i Tabell 4).

Forbinder-felt	Fundament-felt	Nexusfelt			Inholdsfelt		
		V-plads	N-plads subjekt	A-plads	V-plads	N-plads indirekt och direkt objekt	A-plads
Og	*så*	*ville*	*jeg*	*jo ikke*	*have sendt*	*h. bogen*	*tilbage i går.*
Men	*jeg*	*ville*	*[-]*	*jo ikke*	*have sendt*	*h. bogen*	*tilbage i går.*
	Bogen	*ville*	*jeg*	*jo ikke*	*have sendt*	*h. [-]*	*tilbage i går.*
	I går	*ville*	*jeg*	*jo ikke*	*have sendt*	*h. bogen*	*tilbage.*
		Ville	*du*	*ikke nok*	*have sendt*	*h. bogen*	*tilbage i går.*
	Derfor	*sendte*	*jeg*	*jo ikke*		*h. bogen*	*tilbage i går.*
	Så *Så*	*ville* *kom*	*jeg* *skibet*	*jo igår* *jo ikke.*	*have sendt*	*h. bogen*	*tilbage.*
	Så	*kom*	*skibet*			*[-]*	*tilbage i går.*

Tabell 2 Diderichsens huvudsatsschema (aningen förtydligat) efter *Sættningsleddene og deres stilling – tredive år efter* (Diderichsen 1966). *"[-]"* får här beteckna ett 'spår' dvs. markera att ledet på aktuell plats spetsställs till fundamentet. Obligatoriska led enligt denna modell är normalt finit (finita verb) och subjekt (i de flesta huvudsatskonstruktioner).

		N-plads	A-plads	V-plads	V-plads	N-plads	A-plads
og	*om*	*jeg*	*dog ikke*	*ville*	*have sendt*	*h. bogen*	*tilbage i går.*
men	*att*	*jeg*	*nok ikke*	*kom.*			*tilbage i*
og	*at*	*jeg*		*kom*			*går.*

Tabell 3 Diderichsens bisatsschema (aningen förtydligat) efter sidan 371 (1964, s. 371)[2] beskriver en prototypisk bisatsordföljd utan någon motsvarande fundamentposition.

Inledare	Mittfält			Slutfält			
Satsbas (fundament)	Finit verb	Subjekt	Sats-adverbial	Icke-finit verb	Partikel-adverbial	Objekt, eg. subjekt, predikativ och objekt-liknande adverbial	Övrigt adverbial
Ni	*hade*	*[-]*	*nog*	*funnit*	*på*	*något nytt*	*nästa dag.*
Igår	*hade*	*det*	*faktiskt*	*passerat*		*en tankbil*	*på vägen.*

Tabell 4 I *Svenska Akademiens språklära* (Hultman 2003), sida 292, ser det svenska huvudsatsschemat ut på ett inte alltför annorlunda sätt (för enkelhets skull är det lätt modifierat).

Schemat inkorporerar egenskapen som kallas *V2* som finns i nästan alla germanska språk, men som är ovanlig globalt sett, och som innebär att finit verb utgör satsled på position två i deklarativa huvudsatser. Den första positionen, fundamentet, är dock öppet för de flesta andra ledtyper. När ett led inte är flyttat, t.ex. *spetsställt* (placerat i fundamentet, *fundamenterat*) kan det sägas finnas på sin *kanoniska position*.

[2] Ett nordiskt satsschema som på samma gång täcker bi- och huvudsatsmönster har presenterats av Christer Platzack, se t.ex. Platzack (1998), s. 93.

1.2 Satsschemat i parsningsprojekt för svensk text

Medan den positionsgrammatiska beskrivningen med satsschemat har en tungt vägande roll i teoretisk grammatisk analys av svenska måste frågan ställas vilken roll satsschemat har haft i datoriserad syntaxanalys. I stora svenska parsningsprojekt finns satsschemat ofta med, antingen tydligt inkorporerat i grammatikbeskrivningen, eller som en bakgrundsbeskrivning av svensk satsgrammatik. Det följande är en listning av några tongivande system utan att vara en helt uttömmande sammanställning över alla parsningsprojekt för svensk text. Koden är bara tillgänglig i ett fåtal system och satsschemats roll kan därför inte avgöras från enbart beskrivningarna i rapporter.

Satsschemat verkar inte ha använts i två system som troligen är bland de allra första som åtminstone delvis riktar in sig på svensk text, och vilka båda hade ambitionen att på något sätt också göra semantisk analys, nämligen *A Natural Language Parsing Program for Question Answering* (Palme 1971) med ett generellt frassstrukturellt ramverk för bl.a. svenska, eller en kontextfri parser skriven i LISP av Welin (1976). Inget av dessa två antagna tidigaste program verkar ha hög täckningsgrad för fritext.

Satsschemat finns däremot med i senare frasstrukturella projekt under åttiotalet och tidigt nittiotal, t.ex. i *Swedish Syntax* (Ejerhed 1985) och i unifieringsbaserade *Uppsala Chart Parser, UCP* (Sågvall Hein 1987) som båda använder sig av fundamentkonceptet. Bland de svenska parsningsprojekten har Diderichsens satsschema mest framskjuten roll i arbetet i teoretiska *A Grammar Combining Phrase Structure and Field Structure* (Ahrenberg 1990) och i två mindre modellimplementationer. Dessa två använder i likhet med fallet här *inte* en trädstrukturell eller på annat sätt generativ (eventuellt bidirektionell) grammatikbeskrivning enligt vilken textmeningar parsas. Dessa är modellimplementationerna *Nexus Grammar*, *NEXG* (Sigurd och Gawrońska 1994) samt ett arbete av Lindberg och Svensson (1992), se kapitel 5. Båda är Prologimplementationer.[3]

När det gäller mer ytstrukturella, och finite state-betonade parsrar som hanterar fritext så finns satsschemamodellen med, åtminstone som teoretisk modell, i senare utarbetning av både Kokkinakis (2001) och Knutsson (2005). Schemat verkar dock inte ha spelat avgörande roll i utvecklingen av grundimplementationerna av dessa ytstrukturanalysatorer och används inte i den tillika finite state-betonade ansatsen för grammatikkontroll av Sofkova Hashemi (2003). Andra exempel på senare parserbyggen i avhandlingsprojekt är den inkrementiella an-

[3] Ett exempel på en sådan liknande modellimplementation för danska finns i *Danish field grammar in typed prolog* (Rue 1987).

satsen hos Wirén (1992), det unifieringsbaserade systemet av Gambäck (1997) och den datadrivna ansatsen av Megyesi (2002), vilka är tre fall där satsschemat inte heller verkar ha spelat speciellt stor roll. I Källgrens (1989) algoritm som hör till den finite state-influerade *MorP Parser* (Källgren 1992) görs också bl.a. subjekts- och objektsidentifikation utan att satsschemat nämns, även om det kanske kan sägas finnas i bakgrunden.

Helsingfors universitets verktyg *Constraint Grammar – restriktionsgrammatik* (Karlsson, o.a. 1995) rönte stor framgång genom sin metodik med restriktionsregler i valet av analys, och riktade därmed in sig på den ständiga ambiguitetsfrågan, både för taggning och för parsning. Satsschemat finns med som ett rättesnöre i tillhörande litteratur, även om det inte är den definita form vari analysen stöps i versionen för svenska, *SweCG* (Birn 1998). Hur stor roll själva satsschemat får beror på precis hur regelskrivningen görs; formalismen är som sådan flexibel. Detsamma gäller de senare projekt, som inte heller har varit specialiserade just för svenska, och som varit funktionella dependensgrammatiska parsrar. Det första är *SweFDG* av Voutilainen (2001) som överträffar *SweCG* i korrekthet – för en i nämnda källa definierad uppgift. Den andra är dependensgrammatisk parsning *med automatiskt inducerade regler* som behandlar bl.a. svenska, men i princip godtyckligt trädbanksförsett språk: *MaltParser* (Nivre, Hall, o.a. 2007). Det betyder att satsschemat finns närvarande i den mån det gör det i trädbanken, och så kan sägas vara fallet i implementationen för svenska som beskrivs i källan. Vidare finns information från Diderichsens satsschema i svenskaresursen för det typteoretiska ramverket *Grammatical Framework* (Ranta 1994).

Det föreliggande arbetet skiljer ut sig genom att innebära ett systembygge för parsning av svensk text där satsschemat har en viktigare roll än i de flesta, samtidigt som det är öppet för fri text. De nämnda parsrarna för fri text inbegriper en grammatikbeskrivning som är formulerad enligt någon av de olika klasserna i den generativa Chomsky-hierarkin som *kontextfri grammatik, reguljär grammatik*, avknoppningar från dessa (*Head-driven Phrase Structure Grammar, Lexical Functional Grammar* m.fl.) eller andra trädliknande analysformer (*dependensgrammatik*). Rent allmänt kan sägas att satsschemat, rimligt nog, ofta finns i de system som riktar in sig på *funktionell* grammatisk analys, medan flera internationellt influerade system kan ha en rent frasstrukturell analys, kanske eftersom de funktionella satsdelarnas inbördes placering ändå är en mindre fråga i ett språk med mer fixerad ordföljd som engelska.

Skillnaden mellan de nämnda parsningsmetoderna och den modell som beskrivs i kapitel 2 och 3 här, är att analysnivån för resultatet är annorlunda och att ansatsen för parsning ser helt annorlunda ut: Användningen av satsschemat påminner istället här om ordningsföljden för delstegen i nämnda skolövning *primär*

satslösning enligt Tabell 1. Den analys som hittills görs i programmet gäller enbart huvudsatsnivån. Det är dock mycket troligt att denna analys kan utökas till underordnade satsnivåer, med tanke på den mer fixerade ordföljden där blir den totala korrektheten då troligen högre. Att inleda på huvudsatsnivå innebär ett mindre komplext utgångsläge. Flertalet av de ovan nämnda systemen bygger alltså på en formalism som direkt eller indirekt har sitt ursprung i den generativa grammatikteorin. Grammatiken är ofta löstagbar och eftersom den fungerar åt båda håll, både för analys och generering, åtminstone i teorin, kan den kallas bidirektionell. En sådan grammatik är oftast formulerad deklarativt och innebär en språkdefinition, vilket också är en viktig skillnad gentemot det aktuella tillvägagångssättet. I programmet som detta avhandlingsprojekt innehåller finns inte en grammatik som en löstagbar enhet med ambitionen att kunna fungera även för generering, utan systemet är i helhet byggt heuristiskt för just svenska med en procedurell metod formulerad i ett imperativt programmeringsmodus.

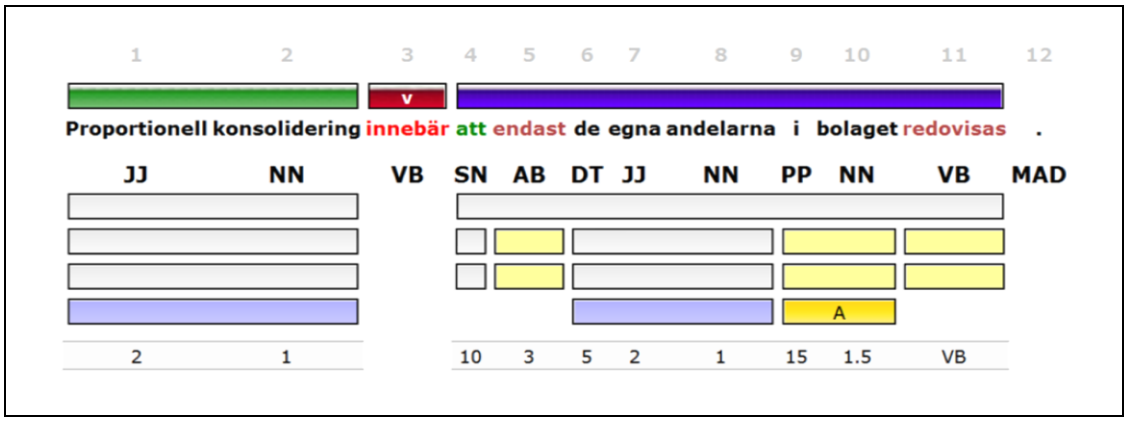
Medan en parsning till en fullständig satslösning av en mening, t.ex. en trädanalys, kan ses som en funktion som har oändligt värdeförråd – dvs. det finns oändligt många olika strukturträd som möjliga analyser – så innebär en satsvis analys med satsschemat i praktiken att resultatformen i hög grad är en återkommande mängd av ledmönster, exempelvis *”adverbial - finit verb - subjekt - adverbial.”* Noga räknat är även dessa möjliga resulterande ledmönster per sats – teoretiskt sett – obegränsade till antalet. Det beror bl.a. på att *adverbial* grammatiskt korrekt kan staplas i princip obegränsat i varje sats utan samordningar. Antalet verb, subjekt och objekt/predikativ är emellertid begränsat per sats och ska vara kompatibelt med verbvalens, modus etc.[4] I kapitel 7 diskuteras vad som skiljer detta tillvägagångssätt för satslösning gentemot vad som kanske kan kallas ’standardmodellen för syntaktisk parsning’. Med detta menas i så fall ett program som innehåller en explicit regelbaserad grammatikkomponent (t.ex. skriven i den kontextfria grammatikklassen) tillsammans med en specialiserad parsningsalgoritm (t.ex. Earleys algoritm).

```
<subjekt>Proportionell konsolidering</subjekt>
<pfv>innebär</pfv>
<objekt>att endast de egna andelarna i bolaget redovisas</objekt>
<tom>.</tom>
```

Kodexempel 1 Den huvudsatsanalys den föreliggande metoden ger genereras bl.a. i ett XML-format. *Pfv: primärt finit verb* (hb09a-051). [5]

[4] Det finns fler konstruktionsaspekter som också teoretiskt sett kan göra antalet led per huvudsats obegränsat, se vidare i kapitel 6.

[5] *Exempelnyckel*: Kod inom parentes anger företrädesvis ID i SUC-korpusen, version 2.0.



Figur 1 Analysen visualiseras i HTML. Det är den övre raden ovan texten som är det faktiska funktionella syntaktiska analysresultatet och som visas med färgkodning, vilken ej framträder i tryck (hb09a-051).

Att det finns ett i praktiken begränsat antal led att identifiera per sats, tillsammans med de tydliga restriktioner som satsschemat sätter upp – t.ex. att precis ett led föregår det finita verbet i huvudsatsen, dvs. i fundamentet, öppnar för en metod som begagnar sig av *uteslutningsmetoden* på ett helt annat sätt än i andra ansatser. En avgörande aspekt av metoden är att den först identifierar de *begränsade* leden på huvudsatsnivå. Det finita verbet har en särställning bland dessa då det är obligatoriskt och har en fast position. Genom att finna det primära finita verbet avgränsas fundamentet, och det kan därmed oftast fastställas utan att dess struktur egentligen behöver analyseras. Vidare innebär identifikation av andra led av begränsade längder (typiskt av längden *ett ord*) på huvudsatsnivå att också de andra fälten gränsas av och ofta ’kapslar in’ satsled av obegränsad längd.

Detta är en av de tänkta fördelarna med schemaparsning. Den komplexa uppgiften att identifiera ett subjekt genom matchning av en flerordssekvens som kan ha en okänd, längd reduceras i många lägen till att avgränsa satsschemats positionsinnehåll med de korta nyckelkomponenterna på huvudsatsnivå. Denna process är implementerad genom en samling speciella regler av omkring några olika grundtyper som *licensierar* (vilket här, praktiskt sett, innebär *att ta bort*) kandidater till de begränsade leden som inte finns på huvudsatsnivå. Metoden innebär alltså att först samla kandidater till de begränsade leden, vilket enkelt kan göras genom ordklasstaggning för t.ex. finita och icke-finita verb. Andra ledslag kräver ordlistor för att identifieras, t.ex. fristående adverbial (*icke, emellertid* m.fl.) etc. Utveckling och förfining av denna licensieringsprocess har pågått under en lång period. Resultatet som visas gäller en korrekthet ifråga om identifikation av dessa avgränsande hörnstenar i huvudsatsen enligt Kapitel 2.

En annan konsekvens av utdataformatet och den analysnivå som detta program hittills ger som resultat är att jämförelser med andra systems korrekthet blir komplicerade och riskerar att bli missvisande. Detta arbete använder huvudsakligen korrekt ordklasstaggad text för undersökning av syntaxanalysens korrekthet, vilket inte gäller i andra projekt. I slutet av Kapitel 3 görs ändå vissa försök att relatera de resultat som visas här till några aktuella parsningsresultat som tidigare har presenterats för svensk text, även om det poängteras att förutsättningarna vid utvärderingar varit alltför olika.

Rent tekniskt innebär detta schemafokuserande program en stor skillnad gentemot t.ex. trädbaserade parsningsprogram som har trädrepresentationen både som ’inre representation’ som byggs med hjälp av en parsningsalgoritm, och som slutresultat. Hur programmet här istället representerar information om en senhet och delresultat under analysens gång beskrivs i Kapitel 4. En djupare diskussion om hur denna sorts syntaxanalys skiljer sig från en som bygger på en modell med mer explicit grammatikbeskrivning finns i Kapitel 7.

1.3 Forskningsfrågor

En vanlig fråga i detta sammanhang är hur väl Diderichsens satsscheman är användbara för modern svensk text, vilket naturligtvis är avgörande. Detta avhandlingsprojekt har inneburit en i grunden positiv syn på gångbarheten. Några vanliga specialfall tas emellertid upp i avsnitt 3.6. För att samma optimistiska hållning ska kunna kvarstå när det handlar om implementerad *parsning* med schemat finns dock flera beroenden. Om det antas att manuell analys Diderichsens satsscheman är möjlig: Är det möjligt, eller rentav fördelaktigt, att som här antas göra *huvudsatsanalys* före en utförlig analys av hela strukturen (motsvarande fullständig satslösning). Och mer precist, eftersom finita verb på huvudsatsnivå utgör de obligatoriska begränsade leden per huvudsats: Hur väl kan dessa absoluta nyckelkomponenter identifieras i svensk text?

Kärnfrågan är alltså om denna metod är möjlig att arbeta efter, för att göra parsning med hög korrekthet, och om det finns syntaktiska fenomen som är omöjliga att lösa. Som redan nämnts överskuggas enkla jämförelser emellertid av det faktum att analysen ser annorlunda ut, jämfört med den hos andra parsningssystem, och kanske därför inte är helt jämförbar. En annan vinkling på samma fråga är som nämnts hur väl en syntaxanalys som inte är formulerad med en explicit regelbaserad grammatik egentligen kan göras.

Här ställs också frågor om vilka typer av applikationer som kan byggas med den aktuella syntaxanalysen. Den aktuella metoden med sin analysnivå antas kunna utgöra en del i många olika språktekniska program. Här presenteras två system-

typer, i kapitel 5, med egenskaper som dock behöver just funktionell syntaxanalys. Det uppkommer fler frågor i kölvattnet av dessa som handlar om automatisk omformulering och informationsextraktion (automatisk frågegenerering till text). Dessa frågor tar över fokus i slutet av detta arbete, även om deras förutsättning är just att själva syntaxanalysen kan ske med tillräckligt bra resultat.

1.4 Representation av svensk text: *Stockholm Umeå Corpus 2.0*

Den grundläggande frågan *vad och hur svensk text är* i ett kvantitativt perspektiv kommer här att ges ett svar genom en representation av svensk publicerad 1990-talstext. I likhet med de flesta andra aktuella större parsningsprojekt för svenska används nämligen den för forskningsändamål fritt tillgängliga genre-uppdelade enmiljonordskorpusen *Stockholm Umeå Corpus* (Ejerhed, Källgren och Wennstedt, o.a. 1992) i detta arbete, närmare SUC 2.0 i SGML-format (Ejerhed, Källgren och Brodda 2006).

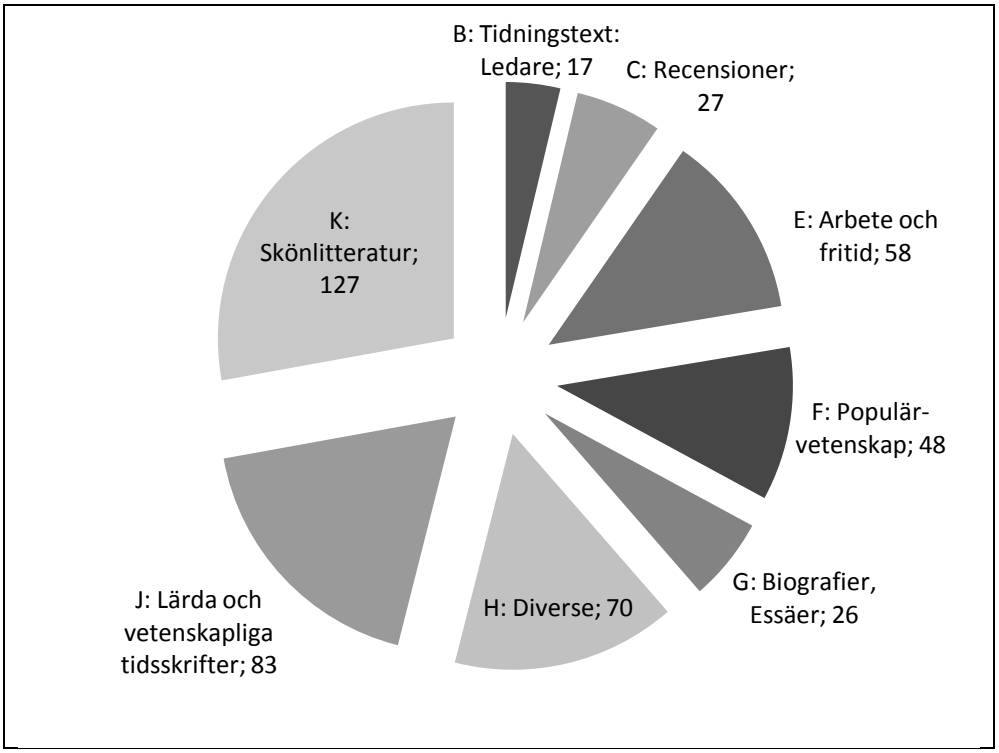


Diagram 1 Antalet filer i de olika huvudkategorierna av text i SUC. Tillsammans utgör de 500 filer. Varje fil tillhör också en mer specificerad underkategori, t.ex. *AC: Ekonomi*.

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 5 Varje löpord i SUC 2.0 är uppmärkt med en av ovanstående ordklasstaggar, oftast kombinerat med särdragsvärden från Tabell 6[6].

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 6 Möjliga särdrag och särdragsvärden i möjliga kombinationer efter *Manual of the Stockholm Umeå Corpus version 2.0* (Ejerhed, Källgren och Brodda 2006). Koder inom parentes innebär att särdragen är tillämpliga på bara en del av ordklassens medlemmar eller att bara en del av särdragsvärdena är möjliga.

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

SUC och de frekvenser för olika grammatiska fenomen som här anges får representera svensk publicerad 1990-talstext, vilket också var avsikten vid dess tillkomst, se Diagram 1 och Tabell 5.[7] SUC 2.0 innehåller svenska texter som publicerats 1990–1994. Varje löpord och tecken är märkt med ordklass och övriga tillämpliga särdrag. Ordet *märkt* förekommer t.ex. med följande två taggningar: *PC-PRF-UTR-SIN-IND-NOM* (particip, perfekt, utrum, singular, obestämd form, nominativ) och *VB-SUP-AKT* (verb, supinum, aktiv). I korpusen är varje löpord uppmärkt med en av ca 150 olika taggkombinationer (ordklass och särdrag) enligt Tabell 6. Genom sin uppdelning i nio huvudgenrer av svensk publicerad text blir resultatet från tester på SUC, kanske mer än någon annan textsamling, talande för vad svensk publicerad text har för egenskaper.

Den ordklass- och särdragsuppmärkning som gjorts i SUC är resultatet av ett övervägande för att ge bra korrekthet i uppmärkningen och samtidigt tillhandahålla en användbar kategoriuppdelning för att t.ex. skriva en parsergrammatik. Somlig information om ord måste dock tillföras generellt senare i en praktiskt tillämpad parser; det gäller hur olika ord med samma taggning måste uppdelas i undergrupper för att de fungerar olika syntaktiskt. Exempel på detta är mängdord och olika adverbtyper (se vidare kapitel 2 och 3, samt appendix). I flera andra system för taggning och parsning, dock ej här, läggs denna information till redan i taggningsprocessen och därmed förändras tagguppsättningen. Se t.ex. Carlberger och Kann (1999) där datum, kopulaverb etc. märks upp eller Forsbom (2008) där olika hjälpverb märkts upp med olika resultat för korrektheten för en därpå grundad ordklasstaggare. I några fall borde taggningen i SUC 2.0 konsekvent kunna ändras. Det gäller t.ex. *respektive,* som oftast taggas som adverb men fungerar som konjunktion, eller *längs,* som alltid taggas som adverb men som nästan uteslutande fungerar som preposition.

```
<s id=ab07c-003>
          <w>Året<ana><ps>NN<m>NEU SIN DEF NOM<b>år</w>
          <w>var<ana><ps>VB<m>PRT AKT<b>vara</w>
          <date>
          <w>1932<ana><ps>RG<m>NOM<b>1932</w>
          </date>
          <d>.<ana><ps>MAD<b>.</d>
</s>
```

Ex 1 *S-enheterna*, här ab07c-003: *(Året var 1932.)*, är de grundläggande enheterna här och har ovanstående utseende i SGML-versionen av SUC, vilken används av det aktuella programmet.

[7] Däremot finns i SUC inslag av 'talspråk' och mycket gammal svenska som ur analysperspektiv har visat sig vara bitvis svåranalyserad. Det är det skrivna ordet som avses med *text* i titeln och generellt i denna avhandling. Därmed inte sagt att satsschemat inte används inom talspråksforskning.

S-enheterna (se Ex 1 och Tabell 7) som alla har sitt eget ID är alltså inte alltid fulla textmeningar med en eller flera huvudsatser, utan även ofullbordade meningar, meningsfragment, rubriker och andra enheter som inte ingår i angränsande s-enheter (se exempel i Tabell 12).[8] I detta avhandlingsprojekt är s-enheterna de grundläggande föremålen för analys som analyseras enskilt en åt gången. Extra informationsmarkeringar i SUC utöver s-enhet och ord med ordklasser, t.ex. datummarkering (*DATE*) eller markering för titlar, beaktas inte speciellt här och läggs ej heller till i själva ordklasstaggarna.

Antal s-enheter	74 157
Antal löpord	1 166 592
Antal unika ord	106 915
Antal olika ordklasser	25
Antal olika ordklasstaggar (ordklass plus särdrag)	153

Tabell 7 Data om Stockholm Umeå Corpus har sammanställts av Megyesi (2002), här i modifierad tabell.

Programmet som är den praktiska implementationen av den här beskrivna metoden (se kapitel 4) har förbättrats manuellt – dvs. regler har skrivits – under en längre period mot en *träningsmängd* av s-enheter. För användning av SUC i detta avhandlingsprojekt har en uppdelning av denna korpus gjorts. Av de drygt 74 000 s-enheterna har 8100 (ca 11 %) slumpmässigt utvalda s-enheter avdelats och använts som *testmängd*. Ett rimligt antagande är att korrektheten för analysen är högre i träningsmängdens s-enheter eftersom dessa undersökts mer och utgör underlaget för regelskrivningen. Därför är träningsmängden använd här när statistik över konstruktionstypers vanlighet i svensk text och frekvenser för olika regelanvändningar klargörs. När *korrektheten* för dessa regelanvändningar

[8] Ett närbesläktat begrepp är *grafisk mening* som används t.ex. i *Mamban* (Teleman 1974) s. 272 och i Melin och Lange (1986). Det innebär en ordsekvens som inleds med stor bokstav och avslutas av ”stort skiljetecken” (punkt, frågetecken, utropstecken eller kolon), men som inte heller behöver vara huvudsatsformad. S-enheterna i SUC blir i jämförelse ett ännu vidare begrepp inkluderande rubriker och andra ”restfragment”.

Ett annat relaterat begrepp är *makrosyntagm* som är den enhet som numreras i *Mamban*. Makrosyntagmer är bl.a. huvudsatser. Mamban räknar med följande typer av makrosyntagmer, både satsformade och icke-satsformade (s. 244): påstående, frågande, imperativisk, utrop, tilltalsfras, interjektionsfras, direkt anföring och rubrik.

Ytterligare ett annat begrepp är det förvillande enkla *mening*. Holm (2000) poängterar hur SAG (Teleman, Hellberg och Andersson 1999) använder det så att uttrycket ”*Först öste han vattnet ur ekan, så rodde han ut på det djupaste stället på sjön, och där lade han ut siknätet.”* räknas som *tre* meningar.

undersökts är det istället analys av testmängden som undersökts. (Se även Figur 30 i Kapitel 4).

Korrektheten för taggningen i SUC 2.0 är den högsta som finns tillgänglig givet storlek och textkvalitet. Det är dock lätt att finna fel och inkonsekvens vid en längre periods undersökning, se avsnitt 4.3. I detta projekt har ansatsen varit att i möjligaste mån utesluta de s-enheter som antas innehålla felaktiga taggningar, både från tränings- och testmängd. När antagen feltaggning påträffats i en s-enhet under utvärderingarna har den så ofta som möjligt också uteslutits, åtminstone om taggningen leder till fel syntaktisk analys med metoden.

De mätresultat som presenteras rörande konstruktioners och konstituenters vanlighet grovt sett, samt den uppmätta korrektheten för analysmetoderna handlar idealiskt om hur vanliga de faktiskt är i svensk publicerad text. Även om det alltid kan ifrågasättas ifall SUC utgör en idealisk representation så är den för närvande det närmsta en sådan som finns. Det som hindrar en alltför viss tolkning om vanlighet är också att frekvensmåtten här, vilka kräver syntaxanalys (t.ex. frekvensen för *primära* verbpartiklar), förutsätter att programmet är helt felfritt i sin uppmärkning. Utöver hantering av denna tillrättalagda textkälla analyseras också fri text som ordklasstaggas av programmet självt. Bl.a. undersöks då text från svenska *Wikipedia*, se Kapitel 5.

1.5 Avhandlingens disposition

Kapitel 2 och 3 är avhandlingens centrala kapitel om syntaxanalys av svenska och beskriver metodik för att finna funktionella led på huvudsatsnivå. Detta sker med hjälp av *licensieringsregler*, en chunkningsteknik för att finna de minsta frasstrukturerna, valensdata och andra manuella regler för att sammanfoga segment på olika grunder. Slutligen görs identifikation av de obegränsade leden genom en kombination av uteslutningsteknik och strukturanalys. Tillsammans med denna metodbeskrivning finns i dessa kapitel studier av förekomstfrekvens för olika grammatiska konstruktioner och konstituenter i svensk text. Dessa mått och listningar av speciella ordgrupper förmodas vara relevanta även för parsning med andra grammatikformalismer och metoder. Kapitel 2 och 3 innehåller en del pseudokod men är huvudsakligen en beskrivning utan att redogöra för programmeringen, så som den faktiskt ser ut. Detta är ett medvetet val eftersom uttryckliga kodexempel risterar att skymma själva metoden. I allmänhet är det också möjligt att programmera de olika delstegen på många olika sätt. I Kapitel 4 redogörs däremot för den programmeringstekniska ansats som valts. Kapitel 5 innehåller beskrivning av tillämpningar av den heuristiska schemaparsningen. Kapitel 6 tar upp framtida forskning och innehåller en teoretisk diskussion.

- *Kapitel 2, Identifikation av primära begränsade satsled,* redogör för metodik för att finna de begränsade leden på huvudsatsnivå, dvs. de positioner i satsschemat som alltid upptas av segment av ett eller ett fåtal ords längd eller är tomma. Detta sker genom licensiering av underordnade ledkandidater efter att bisatser och andra underordnade led identifieras. Denna process urskiljer först finiten, därefter icke-finita verb, adverbial av begränsad längd, partiklar, reflexiva pronomen och ’primära konjunktioner’ vilka samordnar huvudsatser och verbfraser på huvudsatsnivå. Det är kring de i sammanhanget obligatoriska primära finita verben som huvudsatser och primära finita verbfraser bildas. I detta kapitel ingår även beskrivning av identifikationen av förfältsinnehåll, vilka också avgränsar framåt. Förfältsinnehåll är dock egentligen inte alltid segment av begränsad längd.

- *Kapitel 3, Identifikation av primära obegränsade satsled,* redogör för identifikation av subjekt, objekt/predikativ och adverbial på huvudsatsnivå (vilka inbegriper de led som ibland kallas prepositionsobjekt). I de luckor som avgränsats med hjälp av identifikationen av de begränsade leden kan här identifikation ske utan att i samma grad som uppräknade system använda explicit matchning av obegränsade konstituenter. Typfallet är hur fundamentet, enligt satsschemat, bara innehåller ett primärt led. För typbestämning och avgränsning av dessa led från varandra i andra huvudsatsfält beskrivs *rangbaserad chunkning*. Denna procedur skapar chunkar (minimala frassegment). Segmenten utökas därefter stegvis med efterställda attribut för att möta det antal nominala led som verbets valensinformation och satstyp kräver. Identifikation av subjekt och andra led sker här på olika sätt beroende på satskonstruktion.

- *Kapitel 4, Tekniskt utförande*, innehåller en beskrivning av implementationen av den aktuella analysen. Kapitlet innehåller beskrivning av den interna representationen av analysen, aktuell ordklasstaggning och det praktiska förbättringsarbetet. Den praktiska implementationen har både en tillämpad ’visualiseringssida’ och en sida som gör den till en arbetsbänk för kontinuerligt förbättringsarbete av analysen. Programmet är byggt så att ordgrupper och relationer för sammanfogningsregler för segment osv. kan modifieras direkt i gränssnittet när en aktuell analys av en textmening pekar på detta behov. I detta kapitel redogörs även för användning av valensinformation från ordboksdata och en funktionalitet som möjliggör formulering av noggranna sökkriterier, inklusive syntaktiska aspekter, t.ex. för exempelinsamling och analysförbättring.

- *Kapitel 5, Automatisk textvariation samt automatgenerering av besvarade frågor från text*, behandlar två applikationsprototyper för svensk text som

båda använder den aktuella analystypen. Den andra prototypen bygger dessutom på den funktionalitet som beskrivs i den föregående. Att dessa två tillämpningar bygger direkt på analysdelen innebär att ytterligare förbättring av denna gagnar dem direkt. De är också del av samma program och kodbas.

Kapiteldel 5.1 beskriver en prototyp för användarinitierad omformulering av huvudsatser i en ordbehandlingssituation. Denna parafrasfunktionalitet verkar vara den första som hanterar fri text, åtminstone för svenska. Denna prototyp är begränsad till enkla huvudsatser och har främst syftet att demonstrera funktionalitet.

Kapiteldel 5.2 beskriver ett pilotprojekt för automatisk generering av besvarade frågor från deklarativ text. Metoden innebär att från varje möjlig spetsställning i huvudsatsen generera främst *hv-frågor*, dvs. *Idag spelar de* ger teoretiskt sett upphov till *När spelar de? och Vem spelar idag?* Denna implementation är byggd mot svenska *Wikipedias* databas, men tillåter även godtycklig input. I detta system innebär funktionssättet – att uttryckligen generera frågor som har svar – en tydligt annorlunda användarsituation jämfört med andra informationssökningsprogram eftersom bara dessa frågor är tillåtna att ställa. Med andra ord blir det ett system där precision-värdet (andelen korrekta svar för befintliga frågor) speciellt kan gynnas.

- *Kapitel 6, Diskussion och framtida forskning,* innehåller en diskussion om hur den beskrivna analysmetoden kan karakteriseras och hur den kan jämföras med andra metoder för parsning av svenska som innehåller andra typer av grammatikkomponenter. Svar på de generella frågorna som ställts i detta arbete presenteras. Kapitlet berör också vilka delar som kan förbättras givet mer tid för arbetet, och vilka olika möjligheter till förbättring av analys och applikationer som finns. Slutligen tas upp den möjliga forskning som skulle kunna utgå från syntaxanalysen och tillämpningarna som har presenterats.

- *Kapitel 7* är slutligen en sammanfattning på engelska

- *Index* innehåller termer och sidangivelser för var de finns omnämnda i denna avhandling

- I *Appendix* återfinns många av de ordlistningar som använts av programmet och i utvärderingarna.

2 Identifikation av begränsade primära satsled

I detta och nästa kapitel beskrivs metodik för den centrala uppgiften att åstadkomma identifikation av funktionella grammatiska kategorier på huvudsatsnivå mekaniskt, dvs. positionsinnehållen i satsschemat. Det är en beskrivning som framförallt fokuserar på de avgörande stegen för att nå hög korrekthet i svensk publicerad text och i mindre grad på alla de olika grammatiska undantag och ovanligheter som förvisso också ingår i målsättningen.

Fält	För-fält	Initialfält	Mittfält			Slutfält		
Positions-innehåll		Satsbas (fundament) Spetsställt satsled	**Finit verb**	Subjekt	Adverbial	**Icke-finit verb**	Objekt/ predikativ	Adverbial
Struktur/ Längd	Obegr	Obegr	**Begr**	Obegr	Obegr[9]	**Begr**	Obegr	Obegr
	Men	*trots det*	***lycka-des***	*de som startade sist*	*faktiskt*	***vin-na***	*loppet vi nyss nämnde*	*[-] igår.*
Ja, i och för sig –		*han*	***hade***	*[-]*			*ett nytt hus*	*också, fast han inte sa det.*

Tabell 8 Av innehållen i schemapositionerna är bl.a. verben och vissa adverbial begränsade (eng. *bounded*) i längd och struktur. Det innebär att dessa konstituenter är ett, eller ett fåtal ord långa, om de förekommer. Dessa kan tekniskt identifieras genom ordklasstaggning (inklusive grammatiska särdragsvärden) och/eller ordlistning. Det finita verbet har en särställning genom att det är obligatoriskt i huvudsats och fungerar som ett riktmärke för de andra ledens placering.[10]

[9] Även om adverbial på denna position kan ha obegränsad längd och struktur så är det ett användbart faktum att många adverbial som ofta är placerade här, speciellt satsadverbial, är av begränsad längd.

[10] Av s-enheterna i SUC saknar dock mer än 10 % finit på någon nivå över huvud taget. Se vidare Diagram 5.

Som nämnts är *det primära finita verbet,* se t.ex. Tabell 8*,* en konstituent som räknas som obligatorisk, är av begränsad längd (i princip ett ord långt) och har en klar placering (den är oftast nummer två och visar var fundamentet finns). Andra leds närvaro beror på detta led, typiskt uttryckt med syntaktisk valens, medan det omvända inte gäller på samma sätt. Denna konstituent är alltså en uppenbar startpunkt för en fortsatt analys. Den aktuella metodiken tar fasta på att verb (och dessutom verbpartiklar, reflexiva pronomen m.fl., se vidare nedan), är konstituenter med begränsade längder, vilket markerats med fetstil i Tabell 8. Identifikation av dessa led avgränsar i metoden övriga fältinnehåll.

Förutom verben tillhör även verbpartiklar och somliga adverbial (som *ändå* ovan) och konjunktioner dessa avgränsande konstituenter. Genom identifikation av dessa led och schemats information om positioner blir det ofta möjligt att använda uteslutningsmetoden för identifikation av de övriga leden – speciellt i vissa konstruktionstyper som hjälpverbssatser ovan. De begränsade led som identifieras som avgränsande nyckelkomponenter, och som enligt satsschemaperspektivet styckar upp satser och fält när de påträffas är huvudsakligen enligt följande uppställning. Härutöver finns en del specialfall; bl.a. är förfältsinnehåll inklusive s.k. initiala annex, inte strukturellt begränsade men fungerar *avgränsande* i likhet med de andra led som nämns och tas upp i detta kapitel, här med exempel från SUC.

- **Finita verb:** Sedan ***följde*** jag arabens exempel, ***reste*** mig och ***gick*** ut på Aleppos gator. (kk18-119)
- **Icke-finita verb:** Skall ytterligare sprängmassor ***vräkas*** ut? (bb08a-019)
- **Verbpartiklar (här alltid betonade):** Alla föremål strålar ***ut*** värmeenergi. (fh08-035)
- **Reflexiva pronomen:** Sen drog han ***sig*** tillbaks hemma i Älvdalen för att vila på lagrarna. (ae01c-030)
- **Satskonjunktioner:** Nå, jag kom dit ***och*** där satt George Kessler och Vreni. (kk10-136)
- **Fristående (begränsade) adverbial:** Staten satsar ***inte*** pengar på lokala trafikleder. (hb06a-069)[11]
- **Förfält (t.ex. initialt annex):** ***En bjässe till resväska,*** det är sant. (ga05-139)
- **Meningsavgränsande sluttecken:** - Kalla honom inte kung i onödan***!*** (kn09-009)

[11] De adverbial som finns i mittfältet och är av begränsad längd är ofta satsadverbial. Däremot är satsadverbial inte alltid placerade där och inte heller alltid av begränsad längd. Dessa adverbial har *inte* med s.k. *bundna adverbial* att göra.

Med undantag av förfält och meningsavgränsande sluttecken förekommer de nämnda leden både på huvudsatsnivån och på underordnade satsnivåer. Tillsammans med de tre typerna av obegränsade satsled subjekt (inklusive formellt subjekt, när sådant förekommer), objekt/predikativ-kategorin *objekt* (som även inbegriper egentligt subjekt, se nedan) och adverbial som tas upp i nästa kapitel är det dessa konstituenter (fast då förutom satskonjunktioner, förfältsinnehåll och sluttecken) som kallas *satsled*. Även reflexiva pronomen (ordet *sig/sej*) och verbpartiklar ses dock här som egna satsled (ej t.ex. delar av verbet), föranlett av den aktuella metoden och deras möjliga självständighet i satsschemat. Termen *satsled* som förekommer i SAG ”svarar ungefär mot den traditionella grammatikens satsdelar” (band 1, s. 213). I detta arbete motsvarar satsleden nära nog de olika positionsinnehållen i schemat. Brodda (1973) noterar intressant hur termen *satsdel (satsled)* inte har någon bra enkel motsvarighet på engelska (här kan t.ex. användas *functional syntactical constituent* även om denna inte tydligt inbegriper de begränsade leden). Denna översättningsaspekt kan göra att det i brist på bättre namn blir lockande att helt enkelt kalla metoden som åstadkommer funktionell satsledsanalys här för en variant på *shallow parsing* på engelska. Förvisso ingår en sorts segmentering (chunkning) även här som ett delsteg, men namnet *schema parsing* beskriver bättre det egentliga angreppssättet som skiljer sig mycket från t.ex. en finite state-ansats.

Struktur i kapitel 2, Identifikation av begränsade primära satsled

I avsnitt 2.1 visas det skisserade analysförloppet som beskrivs i detta och följande kapitel. Att undvika en uttrycklig grammatikbeskrivning för de led som är obegränsade (dvs. nominala och adverbiella led med frasstrukturmotsvarigheter som t.ex. NP och PP) är en huvudpoäng, men det påpekas att det i stället görs en del annan matchning av flerordskonstituenter som en förutsättning för att kunna identifiera dessa led genom uteslutning.

I avsnitt 2.2 diskuteras formatet för resultatet av analysen. Det görs en jämförelse med den vanligt förekommande analysformen *primär satslösning* (funktionell satslösning på huvudsatsnivå). Det tekniska utdataformatet HTML/XML berörs också.

I avsnitt 2.3 beskrivs *primärfinitsidentifikation*, dvs. automatisk identifikation av huvudsatsers samt med dessa samordnade finita verbfrasers finita verb. Korrektheten med vilken detta kan ske utgör en flaskhals för korrektheten för hela metoden. En samling regler, av ungefär fem grundtyper, för att täcka olika fall av bisatsigenkänning och liknande, används för att heuristiskt nå fram till dessa fasta punkter som beskrivs som obligatoriska i svenska huvudsatser.

Avsnitt 2.4 tar upp identifikation av primära icke-finita verb som i likhet med huvudsatsernas finit är begränsade (närmare bestämt oftast ett ord långa) men som inte är obligatoriska. *Hjälpverbskonstruktioner* är extra fördelaktiga från analyssynpunkt då de kan avgränsa utrymmet (skapa mittfält) för rekursiva konstituenter och underlätta identifikationen av övriga led.

Avsnitt 2.5 beskriver identifikation av andra begränsade satskomponenter (här generellt kallade satsled) på huvudsatsnivå: (begränsade) adverbial, partiklar och reflexivpronomen. Identifikationen av dessa ej obligatoriska led sker på ett liknande sätt som för verben.

Avsnitt 2.6 tar upp vad som här kallas *primära konjunktioner*. Denna term står här för sådana ord och interpunktioner som samordnar huvudsatser och primära finita verbfraser. Med andra ord samordnar primära konjunktioner segment som vardera innehåller ett primärt finit verb. Härigenom avgränsas utrymmet för huvudsatser och primära finita verbfraser som byggs upp kring de primära verben. En annan liknande konstituenttyp som också kan kallas satsled är de som samordnar *icke-finita* verbfraser.

Avsnitt 2.7 beskriver identifikationen av förfältsinnehåll i satsschemat. Dessa är oftast konjunktioner som kan kallas begränsade, men också en del betydligt mer komplicerade uttryck. Även om huvudregeln är att förfält inte innehåller funktionella konstituenter som kan placeras om i satsschemat, vilket gäller för de flesta andra led, är förfältets innehåll inte alls ’funktionslöst’. Det innehåller ofta referenser till och beskrivningar av satsens övriga delar, men kan oftast kvarstå opåverkat före resten då satsen transformeras vid t.ex. spetsställning.

2.1 Heuristisk analys med Diderichsens satsschema – en skiss

Genom att använda de begränsade leden som initialt identifierade avgränsare gör metoden att uttrycklig matchning av de konstituenter som utgör de obegränsade leden och bland annat har NP- eller PP-form inte behöver vara en lika avgörande del av grammatikbeskrivningen. I de parsrar för svenska som använder en explicit språkdefinierande grammatikformalism utgör beskrivningen av dessa strukturer en mycket stor del av grammatiken. Ett exempel på sådana flerordsmatchningsregler finns i Kodexempel 2.

```
np11 →    ADJ (NOB F)* ADJ NOB F? CC (ADJ | ADV-X| NUM) hd=NOB
  |       ADJ ADJ* CC (ADJ | ADV-X| NUM) NOB CC (ADJ | ADV-X| NUM)*
          hd=NOB
  |       ADJ ADJ* NOB (F NOB)+ CC (ADJ | ADV-X| NUM)* (NOB CC ADJ)?
          hd=NOB
  |       ADJ CC ADJ NOB (F NOB)+ F ADJ hd=NOB
  |       ADJ F (ADJ F)* ADJ CC (ADV-X|R0)* ADJ hd=NOUN
  |       ADJ F ADJ CC ADJ ADJ* hd=NOB
  |       ADJ F ADJ NOB F ADV-X? ADJ+ NOB CC hd=NOB
  |       ADJ MSR NOB (F NOB)+ CC (ADJ | ADV-X| NUM)* hd=NOB
  |       ADJ NOB (F ADV? ADJ NOB)+ CC ADJ hd=(NOB|NP)
  |       ADJ NOB (F NOB)+ CC NOB CC hd=NOB
  |       ADJ NOB F ADJ ADJ* NOB (F NOB)+ CC (ADJ | ADV-X| NUM)*
          hd=NOB
  |       ADJ NOB F R0? ADJ NOB (F|CC) (PART|ADJ)+ hd=NOB
  |       ADJ NOB F ADJ NOB ADJ NOB CC ADJ ADJ* hd=NOB
  |       ADJ NOB F? ADJ (NOB F)* ADJ NOB F? CC ADJ hd=NOB
```

Kodexempel 2 Med tekniken *cascaded finite state*-parsning enligt Abney (1997), här exemplifierad med ingående kod från den svenska motsvarigheten *Cass-Swe* (Kokkinakis och Johansson Kokkinakis 1998) identifieras flerordskonstituenter, här NP-strukturer, genom uttryckliga matchningsregler (*Cass-Swe* testar reguljära uttryck så att sökning efter de längsta NP-strukturerna inleder, och kortare och kortare NP-strukturer därefter matchas i de ej markerade återstående segmenten). Detta exempel är en del av NP-beskrivningen på längdnivån 11 ord. Kategorierna svarar mot ordklasstaggar från Parole-korpusen (Järborg och Danielsson 1996), eller är variabler som står för flera taggar. För matchning (identifikation) av NP utan efterställt attribut används över 700 regler.[12]

Förfält	Fundament	Mittfält			Slutfält		
		Finit verb	Subjekt	Adverbial	Icke-finit verb	Objekt/ predikativ	Adverbial
Hur som helst,	*det man ville klargöra och påtalade*	*skulle*	*de som instämde*	*trots att de avsåg annat*	*tillskriva*	*egenskaper som fungerade*	*eftersom de inte kunde vänta.*

Tabell 9 Exemplet visar hur de obegränsade leden – de som uppbär rollerna subjekt, objekt/predikativ och adverbial, och kallas *nominala* respektive *adverbiella* – är (potentiellt) rekursiva (dvs. obegränsade) och kan innehålla eller utgöras av egna satser (bisatser och relativsatser). För att finna de begränsade led som tillhör huvudsatsnivån (speciellt finiten som är obligatoriska) måste därför de finit som finns på underordnad nivå – här understrukna – tas bort för att det ska vara möjligt att använda uteslutningsmetoden.

[12] Dessa regler skulle kunna komprimeras till ett mindre antal men vissa begränsningar i ursprungsprogrammet förhindrade detta. (Källa: Personlig kommunikation med Dimitrios Kokkinakis.)

Det ska nämnas att medan en huvudpoäng i analysen är att undvika identifikation av flerordskonstituenter, som NP, genom uttrycklig matchning genom att möjliggöra denna uteslutningsmetodik i så hög grad som möjligt, så finns matchning av flerordsenheter faktiskt med på andra sätt i analysen. Ett exempel är igenkänning av s.k. koverta komplementerare (vid bl.a. *som*-strykning, se nedan). Att denna flerordsmatchning som är en del av identifikationen av de begränsade huvudsatsleden faktiskt ändå måste utföras har motiverats av hög korrekthet i resultaten av analys med denna strategi. Tabell 9 illustrerar hur de primära begränsade konstituenterna, när de av dessa identifierats, delar upp huvudsatsen på ett förenklande sätt.

2.1.1 Explicit matchning av flerordskonstituenter

Ett genomgående drag för den typ av schemaparsning som här redogörs för är alltså att den fokuserar i mycket mindre grad än flera andra parsningssystem för svenska på att matcha konstituenter som är mer än ett ord långa (över huvud taget *rekursiva konstituenter*[13]). Ansatsen förlitar sig däremot på att genom ettordsmatchning identifiera begränsade led som fungerar som avgränsande hörnstenar.

Det är alltså en central idé att ägna så lite möda som möjligt åt att identifiera den svårfångade gruppen bland leden av obegränsade längder genom explicit matchning och att överhuvudtaget undvika att identifiera flerordskonstituenter genom explicit matchning av ord eller ordklasser i en grammatisk segmentbeskrivning. Istället innebär uteslutningsansatsen här att finna huvudsatsernas fältskiljande nyckelkomponenter av begränsad längd (*begränsade led*) men även *somliga* av flerordskonstituenterna.

Ett första exempel på flerordskonstituenterna som ändå matchas explicit är vissa flerordsadverbial. Detta visar sig vara nödvändigt för att hindra chunkningsprocesserna (se kapitel 3) att foga samman dessa segment med omgivande (fram-

[13] Att en konstituent är rekursiv (dvs. av potentiellt obegränsad längd och struktur) innebär som bekant att den formulerad med omskrivningsregler kan innehålla en kategori av samma typ (t.ex. NP → NP PP), eller kan innehålla en annan rekursiv kategori, t.ex. *PP → Preposition NP* (vilket är en rekursiv definition). Den uppdelning som görs mellan begränsade och obegränsade satsled är just en uppdelning mellan rekursiva och icke-rekursiva led. Rekursion är verktyget som en grammatik med ändligt antal regler behöver för att täcka det oftast oändligt stora språket.

förallt efterföljande) segment. Några exempel på dessa adverbial är *som bäst*, *som sagt*, *i och för sig* och *helt enkelt*.[14]

Ett annat fall av identifikation via explicit matchning av flerordsenheter i den aktuella metoden är som nämnts processen som identifierar *som*-strykning, eller koverta komplementerare (se 2.2.4). En vanlig regel är att det är två angränsande NP-huvudord som markerar en *som*-strykning som i *Det var ett hus [som] jag köpte*. (Det är dock långt ifrån alla angränsande NP-huvudord som faktiskt utgör en sådan inledning av relativsats: *Idag såg jag dem* är ju t.ex. inte ett sådant fall.) Identifikationen av *som*-strykning har en stor samling olika heuristiska matchningsregler (i skrivande stund ca 50 stycken, om än överlappande) som använder sig av explicit matchning av sekvenser av ordklasser, ytterligare särdragsinformation och uttryckliga ord. En viktig hypotes är att det är en lättare uppgift att explicit matcha utvalda flerordssegment som dessa *gränser mellan nominalfraser* än att explicit matcha fullständiga NP-strukturer. Givet hur vanligt förekommande NP-strukturer är, jämfört med t.ex. *som*-strykning som här faktiskt matchas som flerordskonstituenter, möjliggörs, enligt detta antagande, högre korrekthet i slutänden. I denna kalkyl ingår underförstått också hur spridda de olika flerordstyperna är på olika strukturfall.

Ett tredje exempel på fall där matchning av flerordskonstituenter ändå görs är i identifikationen av förfältsinnehåll, dvs. den del av textmeningar som ibland föregår fundamentet. Dessa segment identifieras omväxlande genom uteslutning och matchning. Ett exempel på uteslutning är att sekvensen *”... – [pronomen med subjektskasus] [första primära finita verbet]”* låter delen före tankstrecket bli förfält utan vidare strukturanalys. I andra fall matchas själva förfältet istället, det kan gälla *Eller rättare sagt, Hur som helst* före en rimlig satsbas.

Även om metoden generellt kan uttryckas så att de positionsinnehåll i satsschemat som har begränsade längder först ska identifieras, innebär alltså vissa fall att matchning av flerordskonstituenter används som ett delsteg för att kunna göra detta, alltså just för att säkerställa identifikationen av de begränsade leden på huvudsatsnivå.

2.1.2 Stegen i den funktionella analysen på huvudsatsnivå

Den metod för identifikation av primära led som skisseras här kan sammanfattas på följande sätt. Den går i korthet ut på att finna primära begränsade led, varefter strukturen hos återstående segment i huvudsatsschemat undersöks. Dvs: innehål-

[14] Eftersom dessa inte avslutas med typiska NP-huvudord av rang 1 (se avsnitt 3.3) innebär chunkningsmetoden att segmenten riskerar att inlemma efterkommande ord med t.ex. rang 1.

ler varje sådant segment ett eller flera primära led, är dessa t.ex. NP- eller PP-formade? Med hjälp av huvudverbets aritet/ställighet och direkta följder av Diderichsens satsschema antas det ofta finnas tillräcklig information för att nå långt mot, eller hela vägen till, huvudsatsanalys med denna ansats, förutsatt korrekt ordklasstaggning med särdrag enligt tagguppsättningen i SUC. Detta är en uppfattning som grundar sig på åsynen av ett stort antal textmeningar under en lång periods testande, samt det faktum att den aktuella implementationen har en mycket stor uttryckskraft i den interna representationen av textmeningar (se 4.2). I punktform kan analysen som beskrivs i kapitel 2 och 3 sägas innehålla de följande logiska stegen.

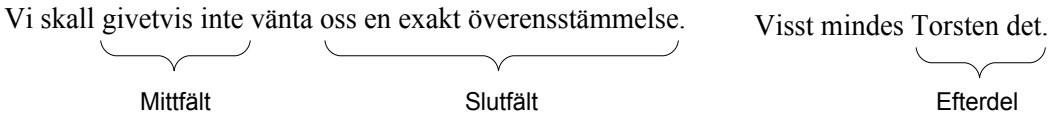
1. *Identifikation av primära finita verb genom licensiering av de icke-primära (detta kapitel)*
   Licensiering kallas det här när verb och senare även andra konstituenter identifieras som syntaktiskt underordnade led om strukturen klargör det. De primära finita verben som återstår ska – om de är fler än ett – kunna motiveras genom huvudsatssamordning eller genom samordning av finita verbfraser (en samordnande konjunktion måste finnas mellan varje par av primära finit). Varje primärt finit verb i en textmening ska vara ensamt i en huvudsats eller primär finit verbfras i samordning. Det är kring de primära finiten, avgränsade med dessa samordnare, som resten av huvudsatser eller primära finita verbfraser ska byggas upp.

2. *Identifikation av övriga begränsade led i varje huvudsats (detta kapitel)*
   Övriga begränsade led som icke-finita verb och fristående ettordsadverbial känns igen direkt genom ordklasstaggning, i kombination med ordlistningar. Det icke-finita verbet är dessutom syntaktiskt beroende av att finitet i aktuell sats är ett potentiellt hjälpverb för att räknas som primärt. Dessa led kan liksom alla andra finnas på både primär och underordnad nivå – därför krävs för alla led av begränsad längd en liknande licensieringsprocedur som för finiten.

3. *Gräns- och strukturanalys och ledidentifikation i de återstående segmenten (Kapitel 3)*
   De primära led som återstår att finna efter identifikationen av de begränsade leden antas oftast kunna identifieras genom en kombination av placering, enkel strukturanalys och valensinformation. Först måste gränser mellan segment av flera obegränsade led konstateras, t.ex. gränsen mellan subjekt och objekt i *Kan den andra personen nya sånger?* Detta steg inleds tekniskt sett med en chunkning, vilket motsvarar gräns- och strukturanalysen. Därefter vidtar ytterligare sammanfogning av segment i flera steg för att skapa de sekvenser som motsvarar de grammatiska kategorierna på huvudsatsnivå. I den praktiska implementationen används här viss data från två valenslexikon.

4. *Identifikation av subjekt, objekt/predikativ och adverbial bland de återstående leden (Kapitel 3)*
   När segmenten i de olika fälten sammanfogats och har strukturbeskrivningar som nominalt/adverbiellt m.m., sker identifikation och etikettering av subjekt, objekt/predikativ och adverbial bland de olika leden av potentiellt obegränsad längd. Metoden för subjektsidentifikationen, vilken har ägnats störst uppmärksamhet av dessa i detta arbete, är uppdelad på flera fall där exempelvis en primär hjälpverbsstruktur innebär speciella förutsättningar (se avsnitt 2.4). Medan antalet nominala led ska svara mot satsens behov av subjekt och objekt/predikativ kan antalet adverbial på förhand inte bestämmas. Just hjälpverbskonstruktion är en speciellt gynnsam konstruktionstyp från ett analysperspektiv, där grundregeln är att ett påträffat nominalt led i mittfältet är subjektet. Om sådant saknas återfinns däremot subjektet per default i fundamentet. (Här finns dock undantag, framförallt i passiva satskonstruktioner). Subjektsidentifikationen här innehåller många undantagsfall där t.ex. snarare verbet avgör tolkningen. En av idékällorna till denna deluppgift finns i Øvrelid (2008). Bestämning av objekt sker senare i syntaxanalysen och med hjälp av en föregående subjektsidentifikation.

Den ovanstående steglistan är en generell beskrivning av tillvägagångssättet för att göra huvudsatsanalys som beskrivs i kapitel 2 och 3. Med hjälp av satsschemat framträder under analysens gång olika ledtrådar som denna ansats tillvaratar. Exempelvis klarläggs fundamentets sträckning ofta redan efter identifikation av det primära finitet, vilket i somliga lägen, kan hjälpa resten av analysen långt framåt.



Figur 2 I detta arbete används termen *efterdel* för att beteckna segmentet efter finitet i en icke-hjälpverbskonstruktion. Mittfält och slutfält är inte avgränsade med ett primärt icke-finit verb i dessa satser och det har visat sig generellt användbart att använda en annorlunda metodik när det inte är fråga om hjälpverbskonstruktion. Rent implementationstekniskt ingår här inte heller de primära verben i fälten utan står mittemellan just som avgränsare. Exempel från SUC: fd02-047 och kk22-139.

Eftersom det är en användbar grundregel att nominala led i *hjälpverbskonstruktioners* mittfält är subjekt och den aktuella ansatsen utnyttjar sådana regler maximalt så testas här till och med en experimentell utvidgning av kategorin hjälpverb, och verbformer med liknande egenskaper, på analystekniska grunder. Denna nytolkning är alltså ämnad att underlätta identifikationen av övriga led. För att tydligt skilja hjälpverbskonstruktioner, som ger speciella förutsättningar,

från andra satser används här en speciell term, *efterdel* (se Figur 2) för hela fältet efter finitet i en icke-hjälpverbskonstruktion.

2.2 Analysnivå och analysformat

En aspekt av parsning kan mycket förenklat beskrivas som en funktion, *parsning:textmening → analys*. Denna funktion *parsning* kommer då att ha olika egenskaper beroende på vilken typ av analys som genomförs. *Värdeförrådet*, dvs. de olika resulterande värdena för *analys*, har egenskaper som varierar beroende på analysnivå. Om det är en fullständig analys, som en full trädgrammatisk analys från en frasstrukturell beskrivning, gäller att det teoretiskt finns oändligt många syntaktiska trädstrukturer i värdeförrådet för ett naturligt språk, liksom oändligt många textmeningar och oändligt många längder på dessa som indata.

Här blir det relevant att fråga hur egenskaperna för parsning till *ledmönster per sats*, t.ex. till *huvudsatsledmönster* fungerar i detta avseende. Eftersom *adverbialled* kan förekomma i obegränsat antal (t.ex. ge ett satsmönster som *subjekt – finit – adv – adv – adv* ...) så innebär även denna analysnivå ett värdeförråd med ett oändligt antal olika analysmönster. Andra strukturer som gör värdeförrådet oändligt stort är bl.a. samordningar av verbfraser i samma satsstruktur, se vidare Kapitel 7. De andra ledtyperna förekommer dock i begränsat antal per sats utan flerledsamordningar: Till varje huvudverb hör eventuellt subjekt, objekt/predikativ och ett begränsat antal reflexivpronomen, partiklar osv. Dessa egenskaper är avgörande, tillsammans med andra förhållanden som att strukturfunktionsrelationen *fristående PP/AdvP – adverbialled* är tydliga (se 3.5.2). Med den form av primärledsanalys som här görs blir de resultatmönster som kommer från analysen *i praktiken* i hög grad en återkommande mängd, som t.ex. mönstret *subjekt – finit – adverbial – objekt*. Det är insikten om att antalet satsled per huvudsats i praktiken är begränsat i kombination med de restriktioner som satsschemat sätter upp som ligger till grund för metoden i projektet.

Att göra en huvudsatsanalys i svensk text kan emellertid innebära olika saker och varierar i specificitet i litteraturen. Den analys som föreliggande implementation gör kan beskrivas som noggrannare än en ren schemaplacering men mindre detaljerad än den form som ibland kallas *primär satslösning*. Jämfört med att bara placera sekvenser ur s-enheter rätt i satsschemat är uppgiften lite mer komplicerad beroende på att *alla* primära satsled helst ska identifieras och avgränsas – alltså även flera primära led (t.ex. objekt eller adverbial) i det som i satsschemat är en och samma schemaposition. Denna åtskillnad ingår i målsättningen, bl.a. för att det är de enskilda leden som i allmänhet kan spetsställas, inte hela positionsinnehållet. Ett exempel på ren schemaanalys utan sådan avgränsning finns i Tabell 10.

Inledare	Mittfält			Slutfält			
Satsbas (fundament)	Finit verb	Subjekt	Satsadverbial	Icke-finit verb	Partikel-adverbial	Objekt, eg. subjekt, predikativ och objekt-liknande adverbial	Övrigt adverbial
Ni	*hade*	*[-]*	*nog ändå*	*kunnat köpa*		*en vän något.*	

Tabell 10 En beskrivning på schemaanalys där gränsen mellan olika primära led inom samma fält inte framgår, dvs. mellan adverbialen *nog* och *ändå,* verben *kunnat* och *köpa* eller objekten *en vän* och *något*.

Det utdataformat som den föreliggande metoden ger har istället form, här formulerat i XML, enligt Kodexempel 3. I likhet med hur satsschemat positionsmässigt hanterar objekt, predikativ och egentligt subjekt som samma positionsinnehåll (N) finns dessa tre konstituenter alla under etiketten *objekt*. I aktuellt arbete och tillämpningar har närmare identifikation av vilken av de tre typerna som döljer sig där inte haft speciell betydelse i analysen. Det har dock antagits vara möjligt att specificera om det rör sig om predikativ eller egentligt subjekt pragmatiskt vid eventuellt behov. I konstruktioner med formellt och egentligt subjekt är det också det formella subjektet som benämns *subjekt*.

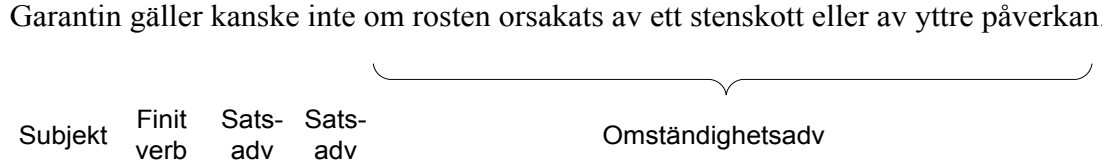
```
<subjekt>Ni</subjekt>
<pfv>hade</pfv>
<adverbial>nog</adverbial>
<adverbial>ändå</adverbial>
<piv>kunnat</piv>
<piv>köpa</piv>
<objekt>en vän</objekt>
<objekt>något</objekt>
<tom>.</tom>
```

Kodexempel 3 Den huvudsatsanalys som ges är mer detaljerad än ren inplacering i schemat som i Tabell 10. Som synes klargörs här gränserna mellan de olika primära satsleden även inom samma ledposition. *Pfv* står för *primärt finit verb*, *piv* står för *primärt icke-finit verb*. Utdataformatet, tekniskt sett, är dels denna XML-form och en HTML-visualisering (se Kapitel 4). Som synes urskiljs varken adverbial- eller objekttyp i utdata.

Analysen innebär i princip att samtliga löpord och skiljetecken i en textmening finns inom en etikett – antingen tillhör varje sådant element något av de gram-

matiska funktionsleden på huvudsatsnivå eller så har det en etikett som 'förfält' eller avslutande sluttecken (som punkt). Detta leder också till frågan om det är korrekt att tilldela varje ord en sådan etikett. I några fall är det svårt att avgöra precis vilket sådant led som ord och tecken tillhör. Det gäller framförallt kommatecken och liknande som finns mellan klara primära led.

Ett närliggande begrepp, *primär satslösning* för svenska, är en analysnivå för huvudsatsanalys som också saknar finkornigheten som en full trädstrukturell analys erbjuder. Beroende på uppfattning om precis vad primär satslösning respektive schemaanalys innebär, kan skillnaden mellan de två uppgifterna variera mycket. Primär satslösning såsom den ofta ser ut är ett noggrannare analysformat än det föreliggande.



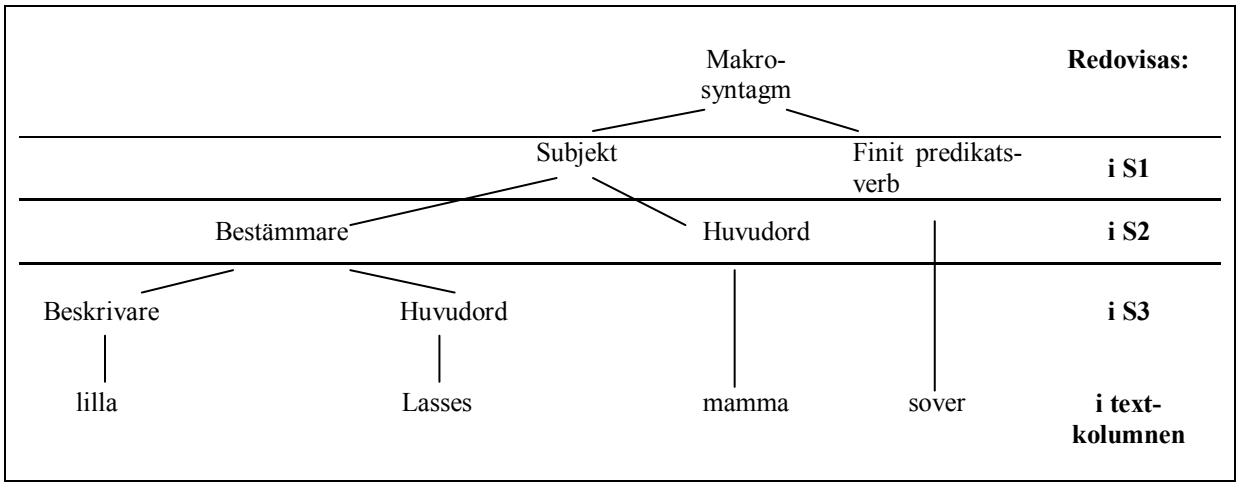
Figur 3 Ett exempel på utförd *primär satslösning* efter Josefsson (2001), övningsbok s. 127. Skillnaden jämfört med den aktuella analysen består i bestämningarna av adverbialtyp. Vidare särskiljs t.ex. direkt och indirekt objekt i förekommande fall.

Primär satslösning av svenska i nyare läroböcker, t.ex. Josefsson (2001), som i Figur 3, innebär jämfört med analysen här att även bl.a. *agent* skiljs ut bland adverbialleden som alla specificeras, t.ex. som tidsadverbial, och är del av uppgiften. I Diderichsens arbete (särskilt Diderichsen 1966) poängterades istället kategorisering av adverbialtyper ifråga om möjliga placeringar. I den aktuella ansatsen urskiljs inledningsvis inte olika adverbialslag åt på dessa grunder, men för omplaceringar blir placeringsrestriktioner viktiga och för frågegenerering blir adverbialtyp (som t.ex. tidsadverbial) också relevant (se Kapitel 5).

Mamba-projekten vid Lunds universitet var tongivande textuppmärkningsprojekt som var kopplade till idéer om kvantitativa syntaktiska måtts relation till texttyp och författare. Analysen i dessa projekt var precis som här funktionellt grammatisk. I Mamban (Teleman 1974) görs en beskrivning av syntaxanalys som gäller samtliga satsnivåer. Målsättningen i föreliggande automatiserade metod är inte olik den översta nivån (den primära nivån, kolumn S1) i Mambaanalysen, men det finns en hel del distinktioner däri, framförallt gällande adverbialslag som inte görs här.[15] I analysen där markeras analysdjupet (ungefär träd-

[15] När det gäller de adverbialtyper som manualen urskiljer antas att dessa skulle kunna införas relativt enkelt med listningar.

djupet) genom placering i olika kolumner enligt nivåerna i Figur 4 och motsvarigheter i Kodexempel 4.[16]



Figur 4 Från ett funktionellt trädgrammatiskt perspektiv placeras de olika nivåerna i kolumner som markerar nivå i Mamban, enligt Kodexempel 4 (Teleman 1974), s. 24.

Text	**S1**	**S2**	**S3**
lilla	Subjekt	Framförställt bestämmarattribut	Framförställt adjektivattribut
Lasses	Subjekt	Framförställt bestämmarattribut	
mamma	Subjekt		
sover	Finit predikatsverb		

Kodexempel 4 Textformatet för analysen i det senare *Mamba*-projektet motsvarar ovanstående trädbeskrivning. Huvudordet är inte markerat i analysen. (Teleman 1974), s. 25. ”Detta betyder att de omedelbara satsdelarna i satsen inte är andra än de som särhålls i t.ex. Diderichsens (Diderichsen 1967, s. 186 (Extern referens)) eller – med vissa modifikationer – i traditionell satslösning” (Teleman 1974), s. 25.

Det ovanstående *Mamba*-exemplet blir med detta arbetes metodik analyserat på huvudsatsnivå, enligt kolumnen `S1`, till utseendet i XML: `<subjekt>lilla Lasses mamma</subjekt><pfv>sover</pfv>`. Identifikation av huvudord sker visserligen även här men har en relativt underordnad roll, och syns för närvarande inte i XML-koden. Även i jämförelse med toppnivån, `S1`, i *Mamba*-

[16] *Mamba*-projekten blir ofta refererade till och syntaxmodellen (som också kallas lundamodellen) har senare använts i minst tre doktorsavhandlingar, nämligen om färöisk text (Sandqvist 1980), barnboksmaterial (Lundqvist 1992) och juridiskt språk (Landqvist 2000).

beskrivningen är aktuell utdata alltså mindre detaljerad. Den är dock nära kompatibel formmässigt och frågan är hur mycket arbete som skulle krävas för att ge just denna form av etikettering i stället. En stor skillnad är t.ex. att vad som i det föreliggande systemet bara är en kategori, adverbial, är uppdelat i många olika grupper. En hypotes är att dessa skulle kunna tillföras med hjälp av ordlistningar. I Appendix återfinns en mappning mellan den mer detaljerade beskrivningen i Mamban och föreliggande format.

Gentemot uppgiften *fullständig syntaxanalys direkt*, fullständig parsning med eller utan grammatiska funktioner, dvs. ’en analys där varje ord har en uttrycklig hierarkisk relation till resten’, kan det skisserade tillvägagångssättet ha några viktiga fördelar: Feltaggning av ord i ett identifierat fundament (så länge det inte handlar om ’viktiga’ ordklasser som finit) har mindre betydelse, eftersom analysen inte måste parsa detta segment uttryckligen till NP eller PP. Denna relativa okänslighet för feltaggning finns också inom andra fält. Metoden är delvis motståndskraftig mot olika typer av strukturell flertydighet inom obegränsade led (exempelvis PP-attachment inom underordnade klargjorda satsled). Förutsättningen för att denna uppgift ska kunna lösas är emellertid att primärfinitsidentifikationen är korrekt. Parsers för fullständig analys av svensk text har också en annorlunda målsättning än vad den aktuella implementationen har, på olika sätt. En relativt ren frasstrukturell utdata produceras som nämnts i Gambäck (1997). Dependensgrammatiska parsrar för svenska som Nivre (2007) och Voutilainen (2001) har fullständig analys (en hierarkisk och funktionsmässig etikett för varje ord) som målsättning och använder syntaktiska funktioner. Därmed innebär en lyckad funktionell dependensgrammatisk analys en noggrannare analys än åtminstone den huvudsatsfokuserande analys som implementerats hittills. I metod är det, som redan påtalats, stor skillnad mellan schemaparsningen och andra beskrivna system. I kapitel 6 görs också några andra distinktioner.

Att frassegmentering (*chunkning,* sammanfogning till frasstrukturer utan efterställda attribut) används i ett delsteg här är en likhet med vad som ibland kallas *shallow parsers*. Schemaparsningens skillnader gentemot shallow parsers såsom *Cass-Swe* (Kokkinakis 2001), och *GTA* (Knutsson 2005) är åtminstone de följande:

- Den aktuella metoden gör en satsnivådistinktion och inleder med primär nivå (huvudsatsnivå).
- Den aktuella metoden har som målsättning att identifiera *hela* primära led, inklusive efterställda attribut, och gör detta i stor utsträckning genom satsschemats angivelser och uteslutningsmetoden.

- Eftersom modellen utgörs av satsschemat och funktionella syntaktiska kategorier som analysnivå återfinns inte vissa frasstrukturella kategorier som *verbfras*.

2.3 Primärfinitsidentifikation

För att åstadkomma automatiserad huvudsatsanalys innebär den metod som beskrivs här att, i likhet med manuell primär satslösning, först koncentrera uppmärksamheten på de finita verben på huvudsatsnivå. De finita verben känns direkt igen genom taggningen, då de är verb i presens, preteritum eller imperativ. Uppgiften är att för varje finit verb avgöra om det är på primär eller på någon underordnad nivå. Denna uppgift som här kommer att kallas *primärfinitsidentifikation* är avgörande för hela analysen på ett drastiskt sätt, eftersom identifikation av de följande primära leden tydligt utgår från att primära finit identifierats korrekt. Arbetet med det har därför pågått under en lång period och det resultat för korrekthet som visas sist kommer från en kombination av de olika delmetoder som beskrivs i detta kapitel. Ändå är det svårt att tro att det från detta läge skulle vara omöjligt att förfina metoden ytterligare. Generellt kan sägas att metoden ger sämre resultat ju fler finit en textmening har – speciellt eftersom delmetoderna för licensiering ibland bygger på utgången från tidigare delmetoder.

Det primära finitet har i huvudsatser i nordiska språk en särställning bland de primära leden genom sin placeringsmässiga stabilitet och att det räknas som ett obligatorium i huvudsats. Närmare bestämt kommer i den tolkning som här görs huvudsatser att kräva ett finit men det gör även samordnade primära finita verbfraser. I Ex 2 är båda finit primära (*är* och *cirkulerar*) och markerade med kursiv stil. Den samordnande primära konjunktionen *och* är understruken.

Ex 2 Vattnet i akvarierna *är* hämtat från fyrtio meters djup i Kosterfjorden och *cirkulerar* ständigt. (aa10b-030)

↑ Primärt finit i huvudsats

↑ Primärt finit i samordnad finit verbfras på huvudsatsnivå

När primära finita verb identifierats betyder det också att eventuella fundament i huvudsatser åtminstone oftast lokaliserats, det är helt enkelt den föregående delen av s-enheten. Även det som i Diderichsens schema kallades *forbinderfelt (förfält)*, i Tabell 1 ovan behöver urskiljas, ifall något sådant förekommer, för att fundamentets sträckning ska kunna säkerställas. Fundamentinnehåll är inte obligatoriskt utan saknas bland annat i *ja/nej*-frågor (*Och kommer de andra?*) och i imperativer (*Stäng fönstret*.). Dessutom hanteras primära finita verbfraser i samordningar, som också har V1-form, på ett liknande sätt som huvudsatserna. I

SUC verkar det vanligaste ordinnehållet i förfält vara konjunktionen *men*. Se vidare avsnitt 2.6.

I de fall en textmening bara innehåller ett finit verb är denna kandidat oftast det primära finitet. Som undantag till detta kan räknas fristående bisatser som är skrivna som meningar med stor bokstav och punkt (*Som om inte det var viktigt.*) där verbet alltså inte är primärt. I SUC har de förekommande s-enheterna (dvs. enheter som oftast är skrivna med stor bokstav och punkt men även rubriker och andra självständiga textenheter som paragrafpunkter och olika fragment) en uppskattad statistisk fördelning med avseende på finitantal som innebär att drygt hälften har noll eller ett förekommande finit.

En slutsats är att primärfinitsidentifikation i ungefär hälften av s-enheterna (om även de s-enheter som helt saknar finita verb räknas) är relativt enkel. Dessa fall utgör emellertid alltså inte alltid självklara fall där det enda existerande finitet alltid är primärt, p.g.a. att s-enheter med bara ett finit alltså ibland är fristående bisatser. Av de återstående s-enheterna är drygt hälften enheter med två finita verb. Enligt beräkningarna utgör s-enheter med högst fem finita verb ca 99,9 % av enheterna i SUC. Att identifiera precis rätt finita verb som primära i meningar med många finit visar sig vara relativt svårt med aktuell metod. Den totala korrektheten påverkas av att delmetoderna har ett visst beroende av varandras korrekthet (typiskt för samordningar).

För att finna finita verb i huvudsatser är det allmänna förhållandet att varje sats har precis ett finit en hjälpsam information. Om en textmening innehåller två finita verb kan dessa båda vara primära om de är samordnade så att varje finit finns i en egen huvudsats eller finit verbfras på huvudsatsnivå. Om ett finit är spetsställt som *vann* i Ex 3 räknas det inte som satsens primära finit.

Ex 3 Vann gjorde SM-ledande Tommy Engvall som var i det närmaste helt överlägsen och fick maskinfel i de heat han inte vann. (ae06f-007)

2.3.1 Termen licensiering

När primära led – först av allt huvudsatsers finita verb – ska urskiljas i en s-enhet med huvudsatsform är det för leden av begränsad längd fråga om att först finna kandidater till dessa, varefter uppgiften är att utesluta de icke-primära konstituenterna bland dessa. Kandidater till finit är som nämnts de ord med taggning

som anger finit verb dvs. verb som i SUC taggats med presens, preteritum eller imperativ.[17]

Metoden för att avgöra vilka finit i en textmening som är primära här är att licensiera de icke-primära, genom att på olika sätt känna igen bisatsstrukturer och andra underordnade strukturer. Eftersom varje huvudsats, även i samordnad sådan, och varje finit verbfras på huvudsatsnivå antas ha precis (eller högst) ett primärt finit, välkomnas dessutom licensiering på lösare grunder i ett slutläge där flera icke-licensierade finit kvarstår i ett segment utan möjlig samordnare (konjunktion eller interpunktion) mellan två finit.

Här ska först tas upp de undantag från den allmänna regeln om ett finit per sats (eller finit verbfras) som finns. Undantag finns främst på bisatsnivå genom strykning av *har/hade* i bisats (*Mannen som [hade] gått återkom*). När det gäller huvudsatsnivå är ett eventuellt exempel på två primära finit i en huvudsats när finit verbfras spetsställs och satsen omskrivs med t.ex. *göra*. Men tolkningen som görs i denna metod är att denna spetsställda del med verb inte är primär som i Ex 3 (se också Formel 4 av Diderichsen).

Inom finite state-tekniker förekommer *licensiering* av verb, dvs. en sorts uttryckligt 'tillåtelse' av verbförekomster genom identifikation av bisatsstarter (inklusive relativsatsstarter), eftersom underordnade satser också oftast kräver finit.[18] Namnet licensiering används emellertid inte alltid när denna process sker. I Hobbs o.a. (1996) används t.ex. inte termen licensiering, men väl tekniken för att täcka relativsatser på engelska i en pseudosyntax är ett exempel på motsvarighet av denna process, enligt Citat 1.

Citat 1 *The material between the end of the subject and the beginning of the main verb group must be read over. There are patterns to accomplish this. [...]*

Formel 3 Subject Relpro {NounGroup | Other}* VerbGroup {NounGroup | Other}* VerbGroup

Formel 3 beskriver hur exemplet *The mayor, <u>who was kidnapped yesterday,</u> was found dead today* matchas genom att relativsatsen får följande beskrivning i matchningsregeln: *Relpro {NounGroup | Other}* VerbGroup {NounGroup | Other}** där relativpronomenet och verbet utgör den minimala konstruktionen (t.ex. i *The mayor <u>who smoked</u> entered*). Noterbart är att denna regel uttrycks i

[17] Det finns somliga konstruktionsmönster där finit inte ska kandidera till primärfinit. I implementationen tas exempelvis inte *vill* i *det vill säga* med. Verb inom parentes (såvida hela meningen inte står inom parentes) är heller inte kandidater.

[18] Ursprungligen kommer dock termen *licensiering* från rent teoretisk syntaxbeskrivning där den används frekvent, t.ex. av SAG och Wasow (1997).

ett delsteg där nomengrupper och verbgrupper redan är identifierade. Med föreliggande metod sker de olika formerna av licensiering i en mindre bearbetad text,[19] och verblicensieringen är förstasteget. Licensieringen av verb som hör till underordnad nivå verkar nämligen kunna ske genom att t.ex. en stackstruktur[20] tillämpas, så att en bisatsinledning innebär att en licensierare (markör för bisatsinledning, *licensor*) hamnar överst på stacken och tas bort när ett finit verb påträffas i en genomlöpning från vänster till höger.[21] Själva licensieringen av verbkandidater (och även av andra begränsade led som senare hanteras på liknande sätt) går delvis till så att en genomlöpning från vänster till höger genom textmeningen samlar på stackstrukturen och lyfter bort en enhet från stackstrukturen om ett finit som inte redan har licensierats påträffas, och samtidigt licensieras då detta verb. Anledningen till varför en sådan stackstruktur kan vara lämplig är det faktum att bisatser kan nästlas som i Ex 4. (Licensierare är understrukna och licensierade finit kursiverade. Primära finit står i fetstil.)

Ex 4 a) Den genetiska orsaken **är** att det arvsanlag som *bestämmer* hur proteinet globin *ska* vara uppbyggt *är* förändrat. (fg03-067)
b) Ganska snart **kommer** vi att märka att relationen mellan vad olika saker *kostar blir* ny. (ec08b-004)
c) Jo, bl.a. att de viktigaste stammarna inom det som *blev* Sveariket *var* svear och götar. (ff01-093)

De följande avsnitten redogör för ett långvarigt arbete med finitlicensiering (för primärfinitsidentifikation) för svensk text som inte verkar ha någon tydlig föregångare i forskningen. Frågan är därefter hur i princip samma licensieringsprocesser kan användas för att utesluta alla icke-primära finit, och övriga begränsade led för svenska. Det antas att alla förekomster av verblicensiering kan täckas in på något sätt av de grundtyper som beskrivs i detta kapitel.

2.3.2 Licensiering genom kända bisatsinledare

Den första delmetoden för att licensiera finita verb är den mest uppenbara. När en känd bisatsinledare (utgående från SUC:s taggning enligt Tabell 11) påträffas är grundregeln att en bisats därmed inleds. Det betyder att det följande tillhöran-

[19] Motsvarighet till *NounGroup* m.fl. finns alltså inte klargjort för detta arbetes metodik i motsvarande skede för licensieringen. Noteras kan också att subjektsidentifikation (som i exemplet snarast motsvarar identifikation av subjektet fram till och med dess huvudord – och inte dess efterställda attribut, relativsatsen) kan göras betydligt enklare och tidigare i engelska p.g.a. den generellt mer fixerade ordföljden jämfört med nordiska språk.
[20] En *stack*-datastruktur är som bekant en behållare där nya element placeras överst (*push*) och eventuell borttagning (*pop*) sker av det element som ligger överst.
[21] Stackstrukturen kan också implementeras som en räknare (en heltalsvariabel) och håller då inte reda på vad som licensierat precis vad.

de finita verbet licensieras och inte längre kandiderar som primärfinit. I SUC förekommer en tagguppsättning där det finns ett någorlunda klart förhållande mellan ordklasstaggar och sådana klara bisatsinledare. (Termen *bisatser* används här även om relativsatser, när ingen uttrycklig skillnad nämns.) I fråga om licensiering hanteras de på samma sätt.

Licensierande ordklasser (bisatsinledare)	Motsvarande SUC-tagg	Exempel
Subjunktion	*SN*	*Eftersom, medan, om*
Frågande/relativt adverb	*HA*	*När, var, hur, som*
Frågande relativ determinerare	*HD*	*Vilken, vilket*
Frågande/relativt pronomen	*HP*	*Som*
Frågande/relativ possessiv	*HS*	*Vars, vems, vilkas*

Tabell 11 De ordklasstaggar i SUC 2.0 som i allmänhet direkt svarar mot bisatsinledning (inklusive relativbisatsinledning).

Befintliga *som* har ofta en komplicerande feltaggning då *som* som konjunktion taggats som bisatsinledare eller vice versa. I SUC 2.0 tillhör *som* de absolut mest feltaggade orden, denna aspekt blir relevant för alla system som utgår från denna korpus eftersom två av de möjliga taggarna (*HA* och *HP*), som i Figur 5, är bisatsinledare medan den tredje (som i Ex 5) istället är konjunktion.

1	2	3	4	5	6	7	8	9	10	11	12
Som	Kalle	hade	sagt	skulle	de	*som*	åt	bygga	en	båt	.
HA	**PM**	**VB**	**VB**	**VB**	**PN**	**HP**	**VB**	**VB**	**DT**	**NN**	**MAD**
	NOM	PRT AKT	SUP AKT	PRT AKT	UTR/NEU PLU DEF SUB		PRT AKT	INF AKT	UTR SIN IND	UTR SIN IND NOM	

Figur 5 Finita verb licensieras (understrukna) p.g.a. föregående typiska overta bisatsinledare (kursiverade), här genom två förekomster av *som*. *Skulle* kvarstår och tolkas som primärt.

Ex 5 *Som* (konjunktion) *målvakt var han fantastisk.*

En licensieringsprocess av en s-enhet innebär alltså en genomlöpning av enheten med följande händelser i sitt enklaste utförande.

- För varje påträffad bisatsinledare: inled en licensieringsprocess. En licensieringsmarkör placeras på stacken.
- Om en licensieringsprocess pågår (dvs. om stacken är icke-tom) när ett finit påträffas så licensieras detta och en licensieringsmarkör tas bort från stacken.

En pseudo-programmeringsnotation för det ovan beskrivna tillvägagångssättet ser ut som i Kodexempel 5. Alla bisatsinledare kallas här `licensor`. Notationen är schematisk, ofullständig och förenklad för överskådlighetens skull. `Number_of_candidates` innebär i exemplet antalet kandidater för rollen som primärt finit efter licensiering.

```
number_of_licensors = 0
number_of_candidates = 0

// Kommentar: den enklaste licensieringsprocessen som använder sig av
// uttryckliga bisatsinledare

For (each word in sentence)
            {
            if (type(word) = licensor)
                        {
                        number_of_licensors++
                        }
            else        if (type(word) = finite_verb)
                                    {
                                    if (number_of_licensors > 0)
                                    number_of_licensors--
                                    }
                                    else
                                    {
                                    add_to_candidates(word)
                                    number_of_candidates++
                                    }
                        else            if (type(word) = non_finite_verb)
                                            {
                                            if (number_of_licensors > 0)
                                            number_of_licensors--
                                            }
            }
```

Kodexempel 5 Pseudokod för grundläggande licensieringsteknik utgående från uttryckliga bisatsinledare (*licensor*) kan implementeras med en stackstruktur (`number_of_licensors`) för att beteckna bisatsnästlingsdjup.

För att ovanstående procedur ska fungera i svensk text finns en rad undantag att beakta. För det första finns undantag i ord med dessa ordklasstaggar som inte alls inleder bisatser. Detta gäller ord av ovannämnda slag i frågor (I *Vad köpte han?* respektive *in situ*-versionen *Han köpte vad?* där *vad* inte är en relativsatsinledare även om den har samma taggning som en relativsatsinledare i SUC (taggningen är HP: *Frågande/relativt pronomen*). Ett annat fall är somliga flerordskonstituenter som, t.ex. i SUC 2.0, är taggade med flera av dessa licensierande taggar innebär inledning av en enda bisats. I Ex 6 visas två konstruktioner, ’snedstrecksalternering’ och ’*vilka-som*’-konstruktion, där två potentiella bisatsmarkörer bara bör räknas som en.

Ex 6 a) *... det segment vilken/vilket gett...* (jc14-107)
b) *Vilka regler som gäller för detta är ännu ej fastställt.* (hf01d-041)

Den första licensieringsprocessen kan alltså beskrivas som en enkel genomlöpning av textsträngen, från vänster till höger, där bisatsinledare gör att sökningen övergår i 'laddat läge' (har icke-tom stack) och för varje sådan kräver ett finit att licensiera. När ett sådant finit påträffas licensieras finitet – det tas bort från de möjliga kandidaterna och en licensierare tas bort från stacken.

Det skulle kunna vara en rimlig hjälp till identifikation av huvudsatsens finit om det kunde antas att varje s-enhet hade åtminstone ett primärt finit verb. Det är dock, som nämnts, så att många är exempelvis fristående bisatser som är avgränsade med punkt eller dylikt. Över en tiondel av s-enheterna i SUC saknar helt finit verb, och bland de som har finit verb är en del fristående bisatser etc., där inget verb är primärt. Bland textmeningarna i SUC finns alltså en stor mängd s-enheter som inte uppfyller satskriteriet p.g.a. att finit saknas. I Tabell 12 visas exempel på några icke satsformade uttryck från SUC. Dessa är alltså s-enheter, enligt SUC, precis som huvudsatserna. Att s-enheter som inte innehåller huvudsats men väl underordnad sats, som vissa i Tabell 12, kategoriseras tillsammans med andra icke-huvudsatser är en likhet med indelningen i olika makrosyntagmtyper som görs i *Manual för analys och beskrivning av makrosyntagmer* av Loman och Jörgensen (1971) där dessa kategoriseras som *meningsfragment* bland ofullbordade meningar (s 34). Eftersom det hör till detta arbetes natur att behandla huvudsatser, eller enheter med primärt finit, får denna grupp enbart en rudimentär strukturanalys och ingen funktionell syntaktisk beskrivning eftersom de inte behandlas av satsschemat. Frekvensuppskattningar som presenteras inkluderar dock dessa s-enheter.

Form	**ID i** SUC	**Exempel**
NP	hb18-004	*Välbevarat fiskekapell.*
PP	kk51-047	*För dessa dagar.*
Infinitivfras	fb02-080	*Att höra ihop*
Vokativ	kk35-057	*Då så!*
Bisats	kk60-182	*Om det provet tas på honom…*
VP	ac03a-008	*Går smärtfritt*

Tabell 12 S-enheter i SUC som inte är huvudsatser har strukturellt sett många olika former. Det sista exemplet ses dock som innehållande ett primärt finit (och adverbial), vilket ändå är en sorts funktionell analys.

Den första frågan som ställs här är: Hur ofta kommer en licensieringsprocedur med de nämnda uttryckliga bisatsinledarna till användning? Det nedanstående är den första av en del grova *frekvensuppskattningar* här. Frekvensuppskattningarna är gjorda med hjälp av den sökfunktionalitet mot SUC 2.0 som beskrivs i kapitel 4. Dessa mätningar sker mot träningsmängden eftersom denna generellt har högre korrekthet (det är träningsmängden som använts för att skriva och kontrollera reglerna). Den större lämpligheten när det gäller frekvensuppskattningar

beror också på att påträffade s-enheter som innehåller antagen feltaggning märkts upp och tagits bort från korpusen och från detta arbetes undersökningar. De flesta frekvensuppskattningarna visar hur ofta analysen markerar ett visst fenomen, vilket betyder att värdet skulle gälla om inga fall missats och om alla markeringar är riktiga (perfekt *precision* respektive *recall*).

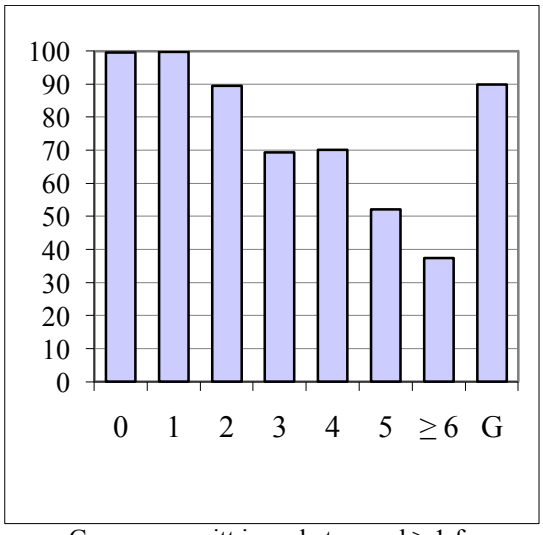
Utvärderingarna av begränsade led som förekommer i detta kapitel rör framför allt korrektheten bland de markerade enheterna (precision). För att snabbt säga något om hur stor andel av olika konstituenter som missats att markera krävs för praktiskt arbete ett manuell trädbanksliknande syntaxuppmärkt facit att utgå från, så detta har inte genomförts. Uppskattning av den totala korrektheten har ändå genomförts genom vissa manuella mätningar som redovisas i kapitlet.

---

15 072 av **40 000,** dvs. **37,68 %** slumpvis analyserade s-enheter var markerade av syntaxanalysatorn som innehållande *minst en uttrycklig bisatsinledare*.

Frekvensuppskattning 1 Förekomst av uttryckliga licensierare enligt beskrivningen ovan innebär att drygt var tredje s-enhet innehåller sådan uttrycklig bisatsstart.

Delmängd	*Full korrekthet*	
S-enheter med 0 fv	188 av 189	99,5 %
S-enheter med 1 fv	629 av 631	99,7 %
S-enheter med 2 fv	366 av 409	89,5 %
S-enheter med 3 fv	127 av 183	69,4 %
S-enheter med 4 fv	40 av 57	70,2 %
S-enheter med 5 fv	12 av 23	52,2 %
S-enheter med ≥ 6 fv	3 av 8	37,5 %
S-enheter inkl 0 fv	**1365 av 1500**	**91,0 %**
S-enheter med ≥ 1 fv	**1177 av 1311**	**89,8 %**



G = genomsnitt i s-enheter med ≥ 1 fv

Diagram 2 Resultatet av primärfinitsidentifikation för en tidig version av enbart enkel licensiering i s-enheter med olika finitantal (*fv*: finita verb) i en undersökning på 1500 enheter visar tydligt bl.a. att satser med många finit är svåra att tolka helt rätt med avseende på finitlicensiering.

Diagram 2 visar resultat från en tidig, oförfinad, undersökning av denna licensieringsprocedur med uttryckliga bisatsinledare (*enkel licensiering*) som relaterar korrektheten till hur många finit de olika s-enheterna har. Observera att som korrekt analyserade räknas här s-enheter i testet vars *samtliga* finit korrekt marke-

rats som primärt eller licensierat. Som resultatet visar kan ca 90 % av korrekt taggade s-enheter ges korrekt tolkning angående finita verbs satsnivå, dvs. huvudsatsnivå eller ej, med denna algoritm.

Det är i implementationen sällsynt med fel i identifikationen av dessa uttryckliga bisatsinledare, förutom i fall av feltaggning. En sökning med en sen version av programmet efter 100 framslumpade s-enheter med minst en markerad uttrycklig bisatsinledare visade att ingen av enheterna hade syntaxanalysfel som berodde på markerad bisatsstart av detta slag. I ett fall markerades en bisatsstart i enlighet med ord och tagg som egentligen inte borde markerats, men som inte fick några konsekvenser.[22]

2.3.3 Licensiering vid identifikation av *som*-strykning och andra strykningar

De nedanstående typerna av licensiering sker på annat sätt än med uttryckliga bisatsinledare. Ex 7 demonstrerar hur relativbisatser kan sakna en typisk bisatsinledare, i fallen nedan: *som*/*vilken*/*då*/*när*. Generellt kommer dessa fall utan bisatsinledare att kallas *som*-strykning (’kovert komplementerare’). Det engelska begreppet *that-deletion* svarar mot detta fenomen som först beskrivs här men även mot ”*att-strykning”* som i Ex 8, vilket kan ge upphov till lite annorlunda strukturmönster som ofta innehåller anföringsverb som primärt finit före ’luckan’, t.ex. *tror* i a).[23] För att genomföra licensiering genom dessa osynliga licensierare måste de identifieras och ’placeras på stacken’ precis som ”synliga” bisatsinledare.

Ex 7
a) […] royalty, t.ex. 5 % på det försäljningspris licenstagaren tar ut. (jc10-078)
b) Det han bedrev var raka motsatsen. (kk77-052)
c) Så fort hon blev stilla kom myggen och knotten. (kl14-035)

Ex 8
a) Vem tror ni jag har i tankarna? (kk03-113)
b) Konflikter förväntade vi oss skulle uppstå framförallt inom […]. (jc11-082)

För *som*-strykningen i Ex 7 finns en tydlig ledtråd i sättet som relativbisatserna fungerar som efterställda attribut till ett oftast substantiviskt huvudord, och eftersom relativbisatser utan särskild bisatsinledare inleds med subjekt (se bisatsschemat, Tabell 3) markeras bisatsinledningen därför ofta av en sekvens av två

[22] Det gäller uttrycket *för vilken gång i ordningen* i ca03d-011: Romanen utspelar sig - för vilken gång i ordningen […].

[23] Den att-strykning som här åsyftas är alltså bara den som i likhet med *som* kan motsvaras av engelska *that*. Det är med andra ord fråga om den bisatsinledande subjunktionen *att*. Termen *att*-strykning är på svenska tvetydig genom den övriga betydelsen ’strykning av infinitivmärke’. Sådan *att*-strykning har ingen påverkan i den aktuella metoden: *Jag kommer (att) vinna?*, *Vi har äran anmoda – vi har äran att anmoda* (Språkrådets hemsida u.d.).

nominalfraser. Ovan utgörs dessa av *försäljningspris licenstagaren* och *Det han* men även av *så fort hon*. Tendensen är att ett sådant segment av två angränsande led (ofta NP) ofta fungerar som licensierare på samma sätt som en uttrycklig sådan. Det har i undersökningar här visat sig att den andra av dessa nominalfraser (relativsatsens subjekt) ofta är just ett pronomen som i flera av exemplen.

Angränsande nominalfraser i sig är ju dock inte alltid ett tecken på relativsatsinledning. I mitt- och efterfältet innebär det t.ex. ofta en övergång mellan två primära satsled av obegränsad längd (subjekt, objekt/predikativ eller adverbial). Att känna igen nominalfraser är, som redan nämnts, metodologiskt ett moment som verkar kräva viss flerordsmatchning, vilket som också nämnts bryter lite mot idén att primärfinitsidentifikation genom licensiering ska ske allra först och oberoende av annat än ord med ordklasstaggning.

Den totala mängden speciella mönster som typiskt markerar en bisatsstart utan *som/att* är svåröverskådlig. Det svåra i själva uppställningen av dessa ligger i att uttrycka mönster som markerar dessa inledningar av relativbisatser utan att av misstag matcha på andra strukturer som angränsande nominalfraser som inte har denna funktion. En generell regel är att ett verb bör följa någonstans efter denna punkt. Det blir emellertid också tydligt efter en tids testande att vissa fasta uttryck (liknande *så fort hon*) generellt döljer en *som*-strykning och inte gärna fungerar på annat sätt i svenska. Arbetet med sådan identifikation är som de flesta andra delar manuellt utfört och grundar sig mest på insamlade grupper av exempel från korpusen. Typfall visas i Tabell 13.

Ex 9 exemplifierar ett sällsynt komplicerat fall av *som*-strykning som är dolt i vad som liknar en regelrätt väl kongruerande NP, (*något av det svåraste jägarlivet*). Det är alltså en variant av det sista fallet i Tabell 13 av ’angränsande nominalfraser’ där genus inte är olika.

Ex 9 Att smyga sig in på en kronhjort är <u>något av det svåraste jägarlivet</u> har att bjuda på. (ea25-125)

Under utvecklingen har igenkänningen av *som*-strykningen i metoden gått från en sorts undantagshantering till en mer betydande del av processen, i takt med att korrektheten ökat och den relativa felandelen p.g.a. icke-identifierad *som*-strykning ökat.

Sekvens av ordklasser/ord (ungefärlig och med exempelord för att bli läsbart)	Konkret exempel
”Angränsande nominalfraser” (*som*-strykning)	
Determinator – substantiv – · NP (subjekt) finit/icke-finit	*Den mark · jordbrukarna lånat*
NP · PNSUB (adv) fv	*saker · vi kan*
NP · denna- NP (adv) handlar	*Bilar · denna säljer*
SUBST-INDEF SUBST-DEF · fv	*musikformer · jugoslaverna utövar*
PRON PRONSUBJ · fv	*är något · vi militärer lär*
Det (genus 1) första (räkneord) · patienten (genus 2) noterar	*Det (genus 1) första (räkneord) patienten (genus 2) noterar (genus 1 ≠ genus 2)*
”Adverbial eller liknande involverat som föregående led” (som-strykning)	
NP bara NP fv	*NP bara hon fick (fv)*[24]
så mycket · NP finit	*så snabbt · vi ville*
Det finit därför · vi	*Det är därför · vi*
Senast · NP finit	*Senast · Ture vann*
”*Att*-strykning med anföringsverb”	
NP anf · NP	*Vi tycker · skatten*
NP anf · krävs	*jag tycker · krävs*

Tabell 13 Några ord- och ordklassekvenser fungerar regelmässigt som *som*- och *att*-strykningsmarkörer. De ovanstående fallen visar grovt (ej exakt) skisserade grundfall som täcks med programmerade regler. I programmet används i skrivande stund ca 50 olika, bitvis överlappande, mönster. I allmänhet gäller att ett mönster fungerar med viss utbytbarhet – bisatssubjekt visas här ofta med personligt pronomen, vilket är typiskt, men andra led som bestämda substantiv är ofta lika klara markörer. Generellt gäller även att adverbial kan följa på detta subjekt (se bisatsschemat) samt att bisatsfinit kan saknas (strykning av *har/hade*). Implementationen innehåller många sådana undantag som stoppar regler från att användas.

Ju desto-konstruktionen ses, när satser hanteras och *som* inte förekommer, som en potentiell *som*-strykningsvariant enligt Ex 10. Språkbruket innehåller en del varianter av denna typ –två *ju*-satser innebär exempelvis ofta att det andra snarast fungerar som *desto* (b). *Ju*-satser behöver inte heller arrangeras tillsammans med ett *desto*-led (c).

[24] En sekvens på den ungefärliga formen *”(NP) bara hon (fv)”* kan motsvara bisatsinledning utan uttrycklig bisatsinledare på fler än ett sätt. Dels förekommer regelrätt *som*-strykning mellan två nominala led (*en maträtt · bara hon gillade*). Sekvensen *Bara det fungerade* där uttrycket har en sorts konditional betydelse, och skapar ett adverbial, bör också identifieras.

Ex 10 a) Fast *ju* högre man *vill*, desto större risker kommer man att ta […] (kk33-055)
b) *Ju* hårdare man *blåser*, ju mer dras papperen ihop. (fg06b-052)
c) Husen var mindre och mindre *ju* längre bak de *stod*, och tillsammans […] (fa03-051)

Ju desto-konstruktion, vare sig den ses som *som*-strykning eller ej (*Ju högre [som] man vill…*), täcks med regler av samma slag som annan *som*-strykning. Här är det relevant att urskilja de *ju desto*-konstruktioner som verkligen samordnar satser från andra som inte ska fungera som licensierare, genom räkning av efterföljande finit (*ju mindre man desto rutigare kostym*).

Strykning av har/hade

En annan strykning får motsatt effekt. Det är en för licensieringsprocessen betydelsefull konstruktionstyp där det istället finns en risk att det egentliga primära finitet blir licensierat. Nedan exemplifieras strykning av *har/hade* i bisats.[25]

Ex 11 a) Inget av de djur som i människans sällskap blivit kosmopoliter *kunde* motstå de frestelser värmen frambragte. (kk11-016)
b) Bostäder och platser som man betraktat som fasta *överges*, människor försvinner. (kk48-031)
c) I USA tycker han att utvecklingen huvudsakligen gått i positiv riktning. (ec08a-084)

I Ex 11 riskerar bisatsinledare (*som, som* respektive *att*) att licensiera primära finit (kursiverade) eftersom *har/hade* saknas i de underordnade satserna. Det icke-finita supinumverbet kan dock sägas signalera att finitet, om det verkligen hade funnits med, redan skulle ha föregått detta icke-finita verb – ett icke-finit verb kan sällan föregå det finita verbet i en bisats. Detta innebär att proceduren kan konstrueras så att ett icke-finit verb ’laddar av’ (tar bort en inledare från stacken) sökandet efter bisatsens finit som relativbisatsinledaren har inlett, dvs. räknaren för bisatsinledare räknas ned med 1. Med andra ord licensierar *som* det *icke-finita* verbet. Tabell 14 är en sammanfattning av dessa nämnda inledande svårigheter för en programmerad licensieringsprocedur.

[25] Strykning av de finita *har/hade* ska ej förväxlas med ha-strykning (alltså den icke-finita verbformen) vilket även förekommer på huvudsatsnivå: *De skulle [ha] ringt.*

Svårighet	Symptom	Betydelse för primärfinitsidentifikation	Möjlig åtgärd
Som-strykning	En bisats finit licensieras inte	En felaktig kandidat stryks ej	Upptäckt av viss sekvens av ord/taggar tyder ibland på bisatsstart
Strykning av *har/hade*	En bisats saknar finitet *har* och kan stryka fel kandidat	Ett senare primärt finit licensieras felaktigt	Avladda licensieringsprocedur på icke-finit verb i supinum

Tabell 14 Konsekvenser av frånvaro av *som* eller *har/hade* i bisats innebär att det för svensk text måste finnas vissa specialregler för att hantera dessa fenomen med en parser.

---

934 av **40 000,** dvs. **ca 2,3 %** slumpvis analyserade s-enheter var markerade av syntaxanalysatorn som innehållande *minst en kovert bisatsinledare*.

Frekvensuppskattning 2 Förekomst av koverta licensierare, *som*-strykning eller *att*-strykning, enligt beskrivningen ovan förekommer och identifieras mycket mer sällan än de uttryckliga bisatsinledarna. De har dock krävt betydligt mer arbete för täckning.

En undersökning av korrektheten i identifikationen av *som*- och *att*-strykning på 100 s-enheter med minst en markering visade att 93 var rätt. 4 var onödiga men ofarliga markeringar och 3 markeringar var fel.

2.3.4 Licensiering genom frågeformade (V1-formade) konditionalbisatser

Ex 12 Ställer man ner kaffekoppen, bör man alltså inte bli alltför förvånad om den faller tvärs igenom bordsskivan! (gb07-032)

Att känna igen V1-formade konditionalbisatser, som i Ex 12, och följaktligen licensiera det inledande finitet är en fråga om flerordsmönstermatchning: det handlar om en V1-formad struktur som varken är imperativ, ja/nej-fråga eller subjektslös 'dagboksstil' (*Vaknade kl åtta, gick ut*.). Ett annat icke-licensierat verb (ibland föregånget av ett adjunktionellt *så*) följer dessutom senare. Reglerna som identifierar dessa strukturer består av kombinationer av boolska värden angående främst ord och ordklasser. Grundregeln för denna identifikation består ungefär av följande programmerade villkorskombination.

- Meningens sista tecken är ej frågetecken
- Det inledande finitet följs ej av partikel eller icke-finit verb
- Det senare finitet som antas vara primärt föregås ej av pronomen med subjektskasus, eller konjunktion (såvida det är ej är adjunktionellt *så* taggat som konjunktion)

- Det inledande finitet får föregås av andra ord om dessa är typiska 'förfältare' (konjunktioner, interjektioner, interpunktioner)

Ovanstående regel ackompanjeras av metodik för att bl.a. även täcka V1-del i passiv diates. I implementationen tas ett trettiotal boolska värden fram för att kombineras till regler som den ovanstående. En konstruktionstyp som är svårare att täcka korrekt är V1-konstruktioner i annan huvudsats än den första i textmeningar. I det fallet sker ett undantag från en i övrigt oemotsagd regel att ett finit som direkt föregås av en satskonjunktion tillhör samma strukturnivå som något av de föregående finiten i textmeningen, dvs. verbfrasen är normalt samordnad med något av de föregående finiten (se vidare *Samordningslicensiering*). Det är inte helt enkelt att uppskatta frekvensen för licensiering med konditional V1 eftersom licensieringen delvis sker i en funktion för heuristisk restlicensiering, se 2.3.7.

---

166 av **40 000,** dvs. **ca 0,4 %** slumpvis analyserade s-enheter i träningsmängden var markerade av syntaxanalysatorn som innehållande licensiering genom konditional V1 i första huvudsatsens fundament

---

Frekvensuppskattning 3 Licensiering genom konditional V1 i första huvudsats markeras i 0,4 % av träningsmängden. Som nämnts identifieras alltså en del av dessa konstruktionsfall, bl.a. de som finns i en annan än första konjunkten, genom en senare heuristisk licensiering, se längre ner. I *träningsmängdens* markeringar här var 144 av 166 markerade s-enheter (86,7 %) helt rätt. 17 (10,2 %) licensierades som konditional V1 fastän de borde licensieras på andra grunder (oftast som anföring), och 5 (3 %) var felaktigt licensierade.

I en undersökning av 38 framslumpade s-enheter från testmängden med markeringar för konditionallicensierng var 34 (ca 89,5 %) rätt, 3 var anföringar och 1 fråga.

2.3.5 Identifikation av anföring

Ex 13
a) - Följ med mi, bad han. (kk25-162)
b) Det fanns tre möjligheter, insåg han. (kl18-074)
c) - Simons värdering av sig själv är nog inte i överensstämmelse med hans lärarinnas, gentog skolinspektörn. (kk25-045)

Att uppgiften som behandlas centralt i dessa kapitel 2 och 3 är identifikation av *huvudsatsens* grammatiska funktioner, tillsammans med dessas fulla sträckning, innebär i somliga s-enheter en kraftig förenkling jämfört med djupare analys. Det gäller exempelvis i anföringssatser som i Ex 13, där långa anförda sekvenser som ibland avslutas med *", säger någon"* i denna ansats tolkas som objekt –

dvs. samma analys som görs av SAG.[26] Den anförda delen, objektet, är ofta satser (med huvudsatsordföljd) men varierar i princip lika friskt som strukturerna hos hela s-enheter. När fri text testas och anföringsverb insamlas uppenbarar sig dessutom en mycket större grupp med denna funktion. Drygt 590 verb (både presens- och preteritumformer räknas för sig, p.g.a. delvis olika funktion) används i skrivande stund (se Appendix).

De nämnda verben används för de mest frekventa typerna av anföringskonstruktioner, där mönstret ”interpunktion/citationstecken finit anföringsverb” (följt av subjekt, anföraren) är absolut vanligast i svensk text. Detta mönster har vad som helst (men något) före, vilket är den anförda delen, och subjekt efteråt (eventuellt adverbial följt av subjekt). I den anförda delen licensieras alla finita verb (och andra led) oavsett struktur. Anföringssatser är ofta komplicerade att analysera av andra skäl. Det gäller framför allt hur en anförd del (objekt) ibland omsluter huvudsatsens anföringsverb med subjekt med mera. Fallen i Ex 14 följs av andra primära led (huvudsats eller primär finit VP) efter subjektet, medan de i Ex 15 låter det anförda ledet fortsätta efter subjektet.

Ex 14 a) - Jodå, säger Jonas, funderar ett tag och konstaterar sedan att han kanske skulle kunna vässa serven och volleyn en aning till tisdagens kvartsfinal. (ae02b-012)

b) - Du har din mission att fylla liksom vi andra, löd svaret, och det blev hon knappast klokare av. (kk04-042)

Ex 15 a) I grunden, säger Robert Broberg, har han ändå alltid vart densamme. (ad02a-116)

b) IDOMENEO, brukar det sägas, är Mozarts tråkigaste opera. (ce01f-009)

c) Fysiskt, säger tränare Raimo Pihl, är Patrik fulländad. (ae03a-019)

”Anföring framåt” med kolon (*Jag sa: Kom hit*) är ovanligt i SUC och har inte täckts hittills i implementationen. (Exemplet tolkas för närvarande i implementationen som två samordnade huvudsatser). En liknande form av syntaktisk underordning av ett segment som potentiellt har eget finit verb och kan vara huvudsatsformat visas i Ex 16, som egentligen inte behandlar anföringar i ovanstående bemärkelse.

Ex 16 a) Eländes elände och Vad var det jag sa, hör till standarduttrycken. (fb01-091)

b) Där var det trevligt att vara polis, är hans sammanfattande omdöme om alla åren. (ec19a-025)

c) Begrav alla vapen och skapa fred och frihet för människorna, är det enkla - och komplicerade - budskapet. (cd02d-011)

[26] I den äldre och nyare *Mamban* görs olika analyser av detta fenomen. Den äldre *Mamban* (Loman och Jörgensen 1971) ser *Hon ska åka till Malmö imorgon, tror jag* som ett objekt med påhängt svagtonigt verb, till skillnad från den nyare (Teleman 1974).

Framförallt b) och c) i Ex 16 visar på risken att ett så komplicerat uttryck som *”, är hans sammanfattande omdöme”,* som är en sorts anföring, kan vara nödvändigt att identifiera genom matchning på något sätt. Dessa typer av underordning där den överordnade satsen inte inleds med ett anföringsverb, eller hjälpverbskonstruktion (*skulle de säga)*, är lyckligtvis en relativt sällsynt konstruktionstyp men den förekommer naturligtvis även hos rena anföringar varför infinita anföringsverb också börjat insamlats för möjliggörande av täckning, om än i blygsammare omfattning. Typen täcks inte bra i nuvarande implementation.

1481 av **40 000,** dvs. **ca 3,7 %** slumpvis analyserade s-enheter var markerade av syntaxanalysatorn som innehållande en anföring enligt grundmönstret som beskrivits.

Frekvensuppskattning 4 Anföring enligt grundmönstret (se Ex 13) markeras i ca 3,7 % av s-enheterna.

En sökning efter 100 framslumpade s-enheter från testmängden med markerad anföring[27] visade att 93 var korrekta, sex var fel som ledde till felanalys och en borde ha inneburit licensiering, fast på andra grunder.

2.3.6 Samordningslicensiering

Ett grundläggande undantag i licensieringsprocessen som inte hittills tagits upp är *finit som direkt föregås av konjunktion*. Detta avsnitt behandlar det faktum att underordnade satser och verbfraser inte bara kan nästlas utan också samordnas. I ett fall som *Eftersom huset som de köpte och sålde var gammalt revs det* finns två bisatsinledare och fyra finit. Eftersom algoritmen som hittills beskrivits skulle licensiera *köpte* respektive *sålde* måste finit som föregås av konjunktion särbehandlas, så att även *var* kan licensieras.[28]

Ex 17 a) Kalle spelade refrängen och sjöng versen.
b) Kalle spelade stycket som han gillade och kunde.
c) Damen Kalle kände och bjöd sjöng.
d) Damen Kalle anställde, motiverade och befordrade åt och drack.

---

[27] Sökuttryck: `meningen.ÄR_ANF==true` (se kapitel 4).

[28] Alla förekomster av ord som är märkta som konjunktioner i SUC fungerar inte så – några undantag är ’parvisa konjunktioner’: *antingen, både, endera, varken, hvarken dels, såväl, hvarken, ömsom, så, blott, dock*. I *Mamban* är dessa, till skillnad från i SUC, inte konjunktioner utan benämns *varslande adverbial*. När dessa samordnar icke rekursiva funktionella led, t.ex. verbfraser (*Han både åt och drack*), är den önskade tolkningen även här den att *både* blir adverbial.

I Ex 17 a) ovan är det tydligt att *sjöng* ska samordnas med *spelade* och följaktligen finnas på primär nivå. I b) är *kunde* däremot samordnat med *gillade* och finns således på sekundär nivå. I likhet med hur licensiering i b) hjälper koordineringsfunktionen att ge *kunde* placering på en underordnad nivå – eftersom finitet där har samma nivåläge som föregående finit – underordnas också *bjöd* i c) idealiskt genom att *kände* först licensierats. Slutsatsen är att koordineringsanalys bör ske *efter* de andra typerna av licensiering som normalt inte får licensiera de finit som föregås av konjunktion. I d) visas slutligen hur kedjan av finit som sammanfogas med konjunktioner (inklusive kommatecken) mellan varje konjunkt bör analyseras. En föregående licensieringsprocess bör licensiera *anställde* men inte *åt*. Koordineringsprocessen bör sedan ske *från vänster till höger* och pågå tills inget efterkommande finit i kedjan är kvar. På detta vis blir *motiverade* och *befordrade* licensierade i tur och ordning eftersom de är samordnade med det först licensierade *anställde*.

Finita verb, i de fall de direkt föregås av konjunktion som samordnar huvudsats eller VP på primär nivå, licensieras inte med den enkla stackbaserade ansatsen. Detta skulle innebära att de konsumerade en licensierare, och bisatsinledarens finit på dess satsnivå kanske därmed inte skulle licensieras. Oftast gäller att ett finit som direkt föregås av konjunktion samordnas med föregående finit och därmed hamnar på samma nivå – dvs. det licensieras om det föregående är licensierat.

Ex 18 a) Utbytet mellan Lund och USA var livligt och Gösta Wennberg som bodde och arbetade i Lund på den tiden tog starka intryck. (ad04a-067)
b) En solbil som är betydligt billigare, men fungerar ändå, är Bertil Sjölinders solkonverterade Fiat 126. (ea07a-069)

I Ex 18 visas exempel på hur finit licensieras med samordningslicensiering. I dessa två fall är tolkningen uppenbar på allmänna grunder – de två finiten *arbetade* respektive *fungerar* skulle inte kunna vara primära givet att de följer på konjunktion efter bisats (när de föregås av sådan konjunktion samordnas de generellt på samma nivå som något tidigare finit).[29] Ex 19 och Ex 20 gäller däremot fall där ett licensierat finit inte direkt föregår.

[29] Det finns dock några slags undantag här, som visar att en textmenings första primära finit kan föregås av konjunktion, om denna inte samordnar med bakåtvarande led som här. Ett exempel är V1-satser som föregås av konjunktion i förfält: *Men sjung då!* Andra exempel är de primära finit som föregås av speciella konjunktioner som inte ”samordnar bakåt”. s.k. adjunktionellt *så,* efter ett adverbial (*Regnar det så stannar jag*) är ett exempel. Dessa är nämligen omväxlande taggade konjunktion respektive adverb i SUC. Även andra första konjunktioner i konjunktionspar (*både, varken* m.fl.) som inte samordnar bakåt utan framåt i satsen är undantag: *Han som spelade både sjöng och lyssnade*. I avsnittet om konditionala V1-satser finns även exempel på att finitet ej samordnas bakåt.

Ex 19 a) Det hindrar markförstöring genom att jorden inte längre sköljs ner i dalen utan fastnar i häckarna. (fc02a-049)
b) För dem båda var läget mycket gynnsamt eftersom Eva lämnade sin åk 3 och skulle ta emot en nybörjargrupp till hösten samtidigt som Elsys åk 1- elever började åk 2. (jb03-022)

Ex 20 a) Erna hörde hur Joel öppnade ytterdörren och hade redan slutat lyssna. (kk74-158)
b) Jag tog för givet att han sov och signalerade till Vilhelm Persson att jag tänkte gå in. (kl17-038)

När sekvensen förekommer på annan plats än i fundamentet är frågan mer komplicerad under givna förhållanden. Ex 19 exemplifierar samordningslicensiering medan Ex 20 visar fall där verbfrasen samordnas på den primära nivån, *hade* och *signalerade* ska alltså inte licensieras. I Ex 20 a) finns ledtråden i det efterföljande adverbialet *redan* som skulle föregått finit i en underordnad satsstruktur, medan *signalerade* i Ex 20 b) egentligen kräver mer av semantik (vilket ändå måste kunna uttryckas med formmedel, inklusive listningar) för att analyseras korrekt.

---

708 av **40 000,** dvs. **ca 1,8 %** slumpvis analyserade s-enheter var markerade av syntaxanalysatorn som innehållande samordningslicensiering

---

Frekvensuppskattning 5 I 1,8 % av s-enheterna används funktionen samordningslicensiering.

I en mängd av 100 framslumpade s-enheter bedömdes 61 vara riktigt behandlade i detta avseende. 34 innehöll något fel, fyra ansågs grammatiskt flertydiga, en var i princip rätt men samordnades med ett finit som licensierats på felaktiga grunder. Hela 19 av de 34 med fel tillhörde kategorin Skönlitteratur i SUC. För närvarande är samordningslicensieringen därmed en svag punkt. Det ska kanske nämnas att det torde vara svårt även med andra ansatser. Givet mer tid här skulle en förbättring här definitivt kunna åstadkommas.[30]

[30] Flera idéer om hur korrektheten skulle kunna förbättras har uppkommit. Det handlar t.ex. om hur primära kopulaverb tenderar att samordnas med andra, och hur diates påverkar.

2.3.7 Heuristisk licensiering av överflödiga finit

De föregående avsnitten har redogjort för de allra vanligaste typerna av bisatsmarkörer, både genom uttryckliga bisatsinledande ord och genom andra syntaktiska strukturer. När reglerna för licensiering inte helt fungerat som tänkt leder det till några strukturmässigt intressanta fall. Grundregeln är alltså att varje huvudsats eller primär finit verbfras ska innehålla precis ett finit, om två icke-licensierade finit finns utan en potentiell samordnare av huvudsatser/finita verbfraser emellan är det ett tecken på att någon ytterligare form av licensierande struktur borde ha upptäckts.[31]

De föregående licensieringsreglerna är formulerade för att inte användas felaktigt utan i säkra fall. Exempelvis krävs för igenkänning av anföringssatser att ett kommatecken eller liknande föregår finit anföringsverb. I ett läge när två icke-licensierade finit kvarstår i samma satsstruktur efter de första reglerna ökar dock sannolikheten för att det är t.ex. anföring utan att textförfattaren har använt komma som är anledning till detta läge, speciellt om det andra finitet är ett anföringsverb och efterföljs av en möjlig subjektsstruktur. Ex 21 är exempelvis ett sådant fall, med två olicensierade finit (*uppträder* och *hoppas*), som matchar denna naiva beskrivning av anföring utan komma, men det är i själva verket en konditional V1-sats i andra konjunkten. Detta steg försöker alltså fånga upp missade fall och licensierar också verb. Detta sker dock genom generösare regler än för de redan behandlade typerna av licensiering.

Ex 21 - Det finns inget recept, men uppträder man korrekt hoppas jag att det smittar av sig. (eb06c-022)

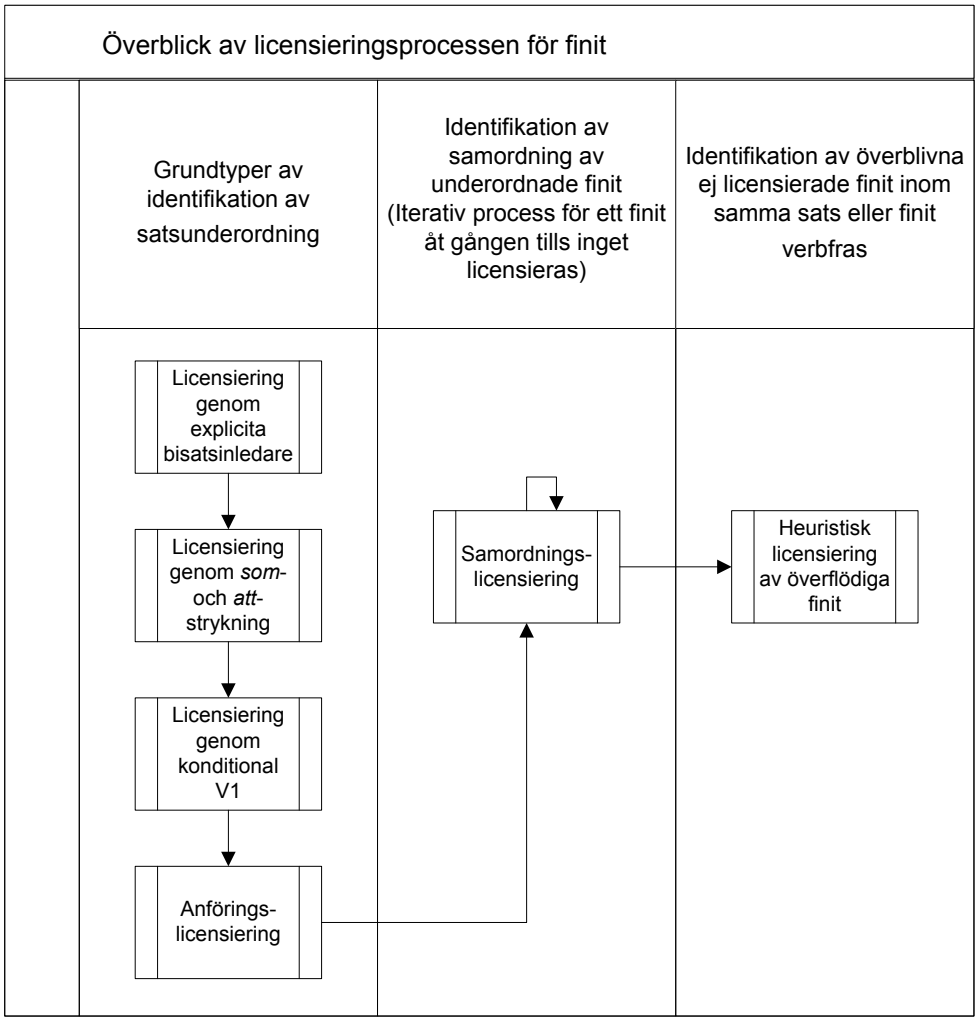
Denna heuristiska licensiering sker med en frekvens som varierar under utvecklingen och med en varierande korrekthet, ibland har den skett i ca 1 procent av s-enheterna. En speciell orsak till detta är att reglerna egentligen gäller de övriga nämnda licensieringsteknikerna. Om en idé om varför licensiering bör genomföras uppkommer så implementeras den i detta arbete ofta direkt i de ovan nämnda ’grundfunktionerna’ för licensiering.

---

[31] I systembyggen, mekaniska som programtekniska, talas ibland om skillnaden mellan *återkoppling* (*feedback*) eller *öppen styrning*. Ett system med återkoppling ger en sorts aktuellt resultatvärde tillbaka till den aktuella funktionen. Ett system som enbart har öppen styrning (t.ex. en elektrisk spis utan termostat) känner däremot inte till det aktuella utdatatillståndet och funktionaliteten är negativt påverkbar av yttre faktorer. I detta arbete sker återkoppling speciellt tydligt i kontrollfunktionen som räknar icke-licensierande finit per möjlig konjunkt. Det är dock möjligt att arbeta fram fler återkopplingsmetoder: De syntaktiska felanalyser som görs ger nämligen ofta upphov till märkliga och ibland omöjliga analysstrukturer, t.ex. första primära finit, någonstans föregånget av licensierat finit, där det antagna primära finitet är direkt föregånget av normal (ej parvis) konjunktion.

2.3.8 Samordning av delmetoderna för licensiering

De olika delmetoderna används delvis beroende på resultatet från föregående delmetoder och körs beroende på antal kvarvarande icke-licensierade finit vid olika punkter i licensieringsproceduren, ordningsmässigt enligt Figur 6.



Figur 6 Licensieringen av finita verb kan ses som tre delsteg. Det är kring de kvarstående finita verben på huvudsatsnivå som programmet bygger upp huvudsatser och primära finita verbfraser.

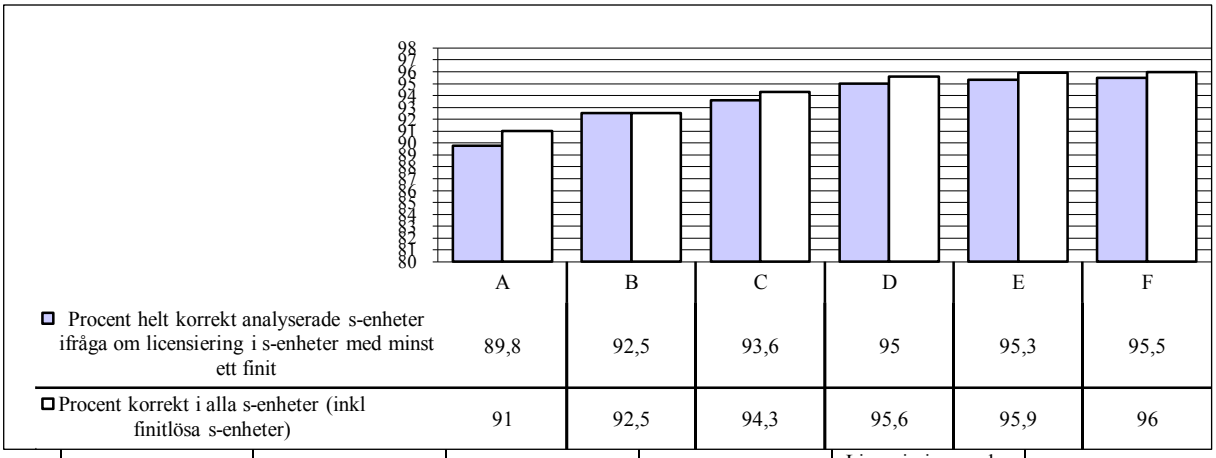
2.3.9 Testresultat för primärfinitsidentikation genom licensiering

Eftersom primärfinitsidentifikationen är en så central del i den parsningsmetod som beskrivs har den omarbetats och utvärderats flera gånger under utvecklingen. I detta avsnitt redovisas först ett tidigt testresultat som motiverar de olika delmetoderna genom bidrag till korrektheten hos hela processen. Härefter, som avslutning, följer slutliga manuellt utförda beräkningar av resultat från den samlade processen.

Det nedanstående resultatet från ett tidigt stadium visar att de olika metoderna alla bidrar till korrektheten. Dessutom ger diagrammen en allmän bild av ungefär vilken korrekthet som kan väntas om texten är korrekt ordklasstaggad. Det är en fråga om minst ca 90 % korrekt analyserade textmeningar, och vid tillfället för denna tidiga resultatuppskattning upp mot 96 % (jämför med den slutliga utvärderingen längre fram). De aktuella delmetoderna var licensiering med uttryckliga bisatsinledare, identifikation av *som*- och *att*-strykning, anföringskonstruktion, V1-formade konditionalbisatser, samordningslicensiering och heuristisk licensiering av ett av flera finit som annars skulle finnas i samma sats och på huvudsatsnivå.

Resultaten som först visas här i Diagram 3 och Diagram 4 kommer alltså från en tidig utvecklingsversion av implementationen, med just syftet att visa att alla delmetoderna hade betydelse. Ett annat syfte var vid tidpunkten för denna tidiga mätning att undersöka vilka textmeningar (med vilket antal finit) som stod för de flesta felaktiga analyserna. Det var då uppenbart att längre meningar med större antal finit gav sämre resultat, men å andra sidan är textmeningar med riktigt många finit, som visats, relativt ovanliga.

Det fanns i detta test ingen undersökning av korrektheten för primärfinitsidentifikation för fallet då ingen delmetod alls användes (dvs. vad korrektheten skulle bli om alla finit antogs vara primära) men licensiering med uttryckliga bisatsinledare är däremot hela tiden med. De olika delmetoderna har, som redan nämnts, beroenden mellan sig – det gäller t.ex. för samordningslicensiering som bör ske när ’normal’ licensiering först genomförts av finit med vilka andra finit alltså samordnas. För detta tidiga test sammanställdes en testmängd på 1500 slumpmässigt utvalda s-enheter från SUC 2.0 i vilka primära finit märktes upp för att möjliggöra snabb återkommande testning där de olika delmetoderna lades till stegvis. För varje s-enhet sparades även det totala antalet finita verb i s-enheten – för att kunna visa hur korrektheten varierade med olika antal finita verb. När uppenbar feltaggning i SUC, som skulle göra korrekt licensiering omöjlig, upptäcktes, så togs dessa s-enheter bort från testmängden. För att räknas som en korrekt analys krävdes här att varje finit i en enhet korrekt markerats som primärt eller icke-primärt (licensierat). Här har de olika delmetoderna lagts på stegvis för att åskådliggöra den relativa förbättring de bidrar med. De har inte utvärderats helt enskilt.



	Enkel licensiering	Licensiering genom igenkänning av anförande	Licensiering genom samordning	Licensiering genom *som*-strykning	Licensiering med frågeformade konditionala bisatser	Heuristisk licensiering
A:	X					
B:	X	X				
C:	X	X	X			
D:	X	X	X	X		
E:	X	X	X	X	X	
F:	X	X	X	X	X	X

Diagram 3 Resultat av en tidig version av finitlicensiering visas här. Resultatet kom från en automatisk testning på 1500 slumpmässigt utvalda s-enheter från SUC 2.0. De olika delmetoderna bidrog här till en sammanlagd korrekthet i över 95 % av s-enheterna (96 % om även finitlösa enheter räknades och då alltid var korrekta). *Licensiering genom igenkänning av närstående finita verb* innebar en tidig variant av licensiering av övriga fall enligt ovan (’heuristisk licensiering’).

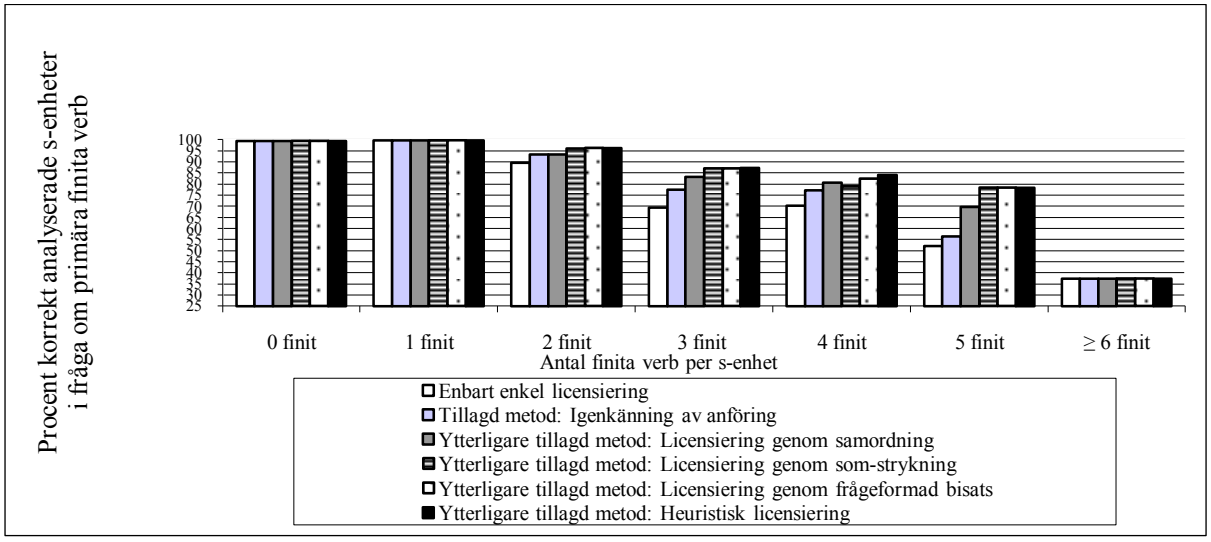


Diagram 4 Samma resultat är här utslaget på korrekthetsprocent hos s-enheter med olika finitantal. (Se längre fram för slutliga resultat.) Som åskådliggjordes i föregående tabell ökade varje delmetod *den totala korrektheten*.

Den totala korrektheten för primärfinitsidentifikation

En slutlig undersökning av korrektheten för primärfinitsidentifikation överhuvudtaget gjordes i slutskedet av detta arbete. S-enheter framslumpades ur testmängden tills antalet korrekta markeringar med avseende på primära finit i huvudsats och primär finit verbfras uppgick till 200 s-enheter. Det innebar 203 huvudsatser/primära finita verbfraser: ca 98,5 var rätt. De tre felen bestod av: ett av misstag samordningslicensierat finit, en anförd del sent i en anföringssats (som därmed borde vara del av objektet) och en annan anföring som missats eftersom *halvviskar* ej fanns bland listningen av hjälpverb.

Hur vanligt är det med olika finitantal och i vilken grad är de primära och licensierade?

Generella sammanfattande frågor om finit och licensiering är a) hur ofta finita verb förekommer och b) i vilken grad de då är primära eller licensierade. Det nedanstående är resultatet i diagramform av en undersökning i en framslumpad mängd av 2011 enheter där det undersöktes hur vanligt förekommande enheter med olika antal licensierade finit, primära finit, och med finit över huvud taget var. Som påpekas innebär det inte så stora testunderlaget här uppskattningar. Denna undersökning gjordes på en relativt sen version av de programmerade licensieringsprocesserna varför de olika andelarna troligen gäller relativt väl för den version av programmet som fanns i slutskedet av avhandlingsarbetet.

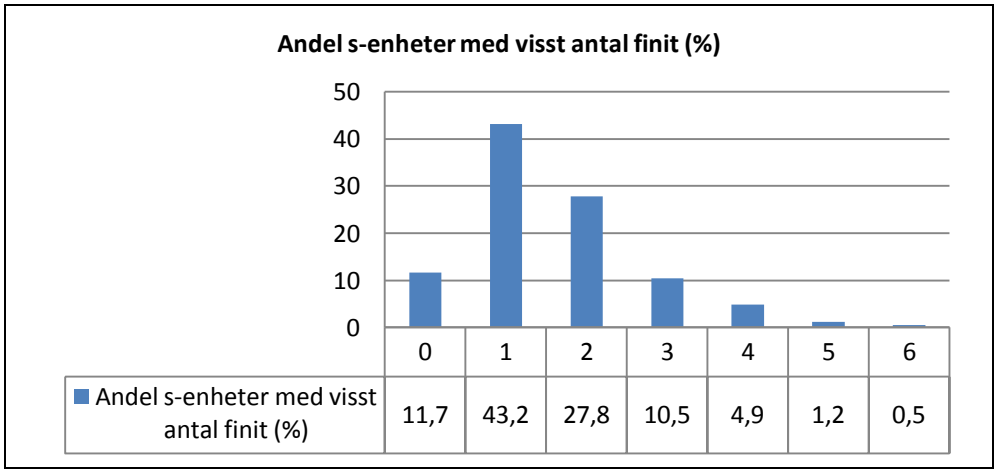


Diagram 5 Andel av s-enheter med olika antal finit (oberoende av satsnivå) visar att drygt hälften hade noll eller ett finit verb. (Punkt 6 innebär 6 eller fler verb.)

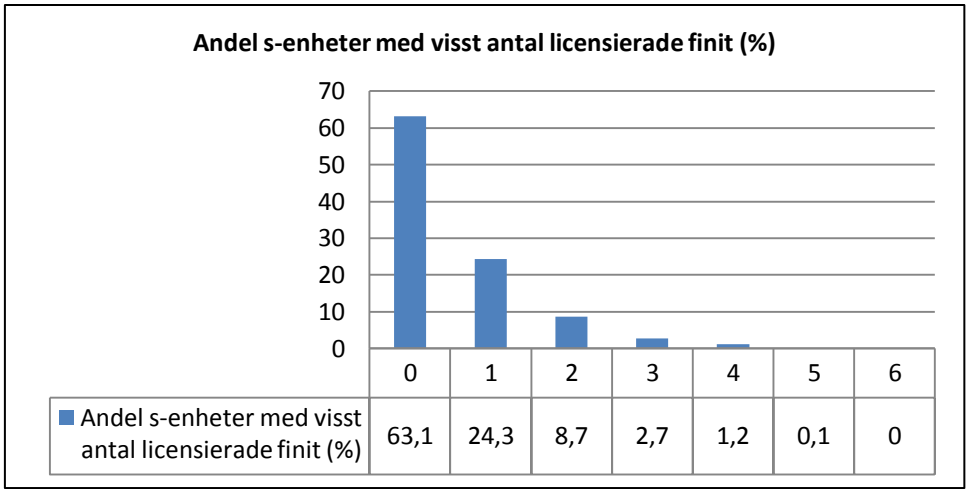


Diagram 6 Uppskattningsvis 37 % av s-enheterna innehöll finit som licensieras. Givet att alla s-enheter inklusive de utan finitförekomster räknas in i detta arbetes mätning så är dessa värden i överensstämmelse med t.ex. Melin och Lange (1986) som i en undersökning på blandad text (s 169 i källan) anger ungefär en bisats per två makrosyntagmer.

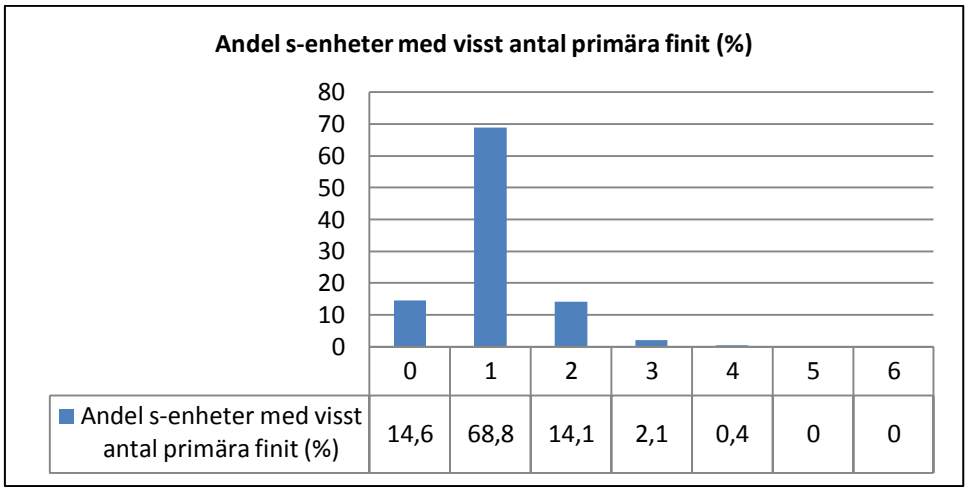


Diagram 7 Antalet primära finit (dvs. kvarstående finit efter licensiering) var ett i ca 69 % av s-enheterna.

2.4 Identifikation av primära icke-finita verb

Precis som alla andra funktionella led i satsschemat finns representanter av icke-finita verb på underordnade nivåer. Det innebär att det kan vara användbart att använda licensieringsidén också för att utesluta icke-finita verb på icke-primär nivå. De många regler som formulerats för licensiering av finita verb används nämligen också för alla andra begränsade led. För andra begränsade led gäller också att dessa kan licensieras, förutom genom identifikation av underordnad satsstart, genom identifikation av satsförkortningstypen infinitivfras. Det innebär

exempelvis att verbpartikeln *till* i *De gillade att lägga till vid bryggan* inte kan bli primär eftersom den ingår i en infinitivfras.

Innan beskrivning sker av det första övriga begränsade ledet, icke-finita verb, ska dock först behandlas vilket förbättrat analysläge för hela satsen som förekomsten av primärt icke-finit verb, och därmed en hjälpverbskonstruktion, innebär. Här måste framförallt klargöras vad som ska räknas som en hjälpverbskonstruktion.

2.4.1 Följden av primära icke-finit: primär hjälpverbskonstruktion

I de fall då en huvudsats efter licensieringsprocesser innehåller primärt icke-finit verb på annan plats än i fundamentet är huvudsatsen av typ hjälpverbskonstruktion, vilket innebär speciellt goda möjligheter för enkel identifiering inklusive etikettering av de kvarstående leden. Antalet subjekt och objektiv/predikativled är begränsat och för varje säker identifikation blir identifikation av resten enklare, enligt uteslutningsmetoden. När fält tydligt avgränsas genom att de begränsade leden identifieras, begränsas vidare möjligheten för olika konstruktionstyper i enlighet med satsschemat (fundamentet rymmer i princip bara ett primärt led, mittfältet i en hjälpverbskonstruktion rymmer oftast bara ett nominalt led). Diderichsen (1966), s. 369, beskrev med bokstäver – som här svarar mot *led* snarare än *positioner* – sin formel,[32] utgående från sin närmast korpuslingvistiska beskrivning i Citat 2.

Citat 2 *Hvis man går ud fra tredelningen i nominale, adverbielle og verbale led, (betegnet N, A, V) og samler en betydelig mængde hovedsættningar fra moderne dansk normalprosa der indeholder mindst to led af hver type, vil man se at de på første plads kan have led af enhver af de tre typer, men at man derefter som regel finder typerne ordnet i to grupper, der hver indeholder mindst et led af hver type, og som bægge har leddene ordnet i samma rækkefølge.*

Formel 4
A
N / V N A.. / V.. N.. A..
(V)

I Formel 4 från samma text betecknar två punkter att flera led av samma typ kan stå efter varandra, medan snedstrecken är fältgränserna. Formel 4 är en koncis sammanfattning av den bedrift den schematiska beskrivningen innebär: att ha

[32] Till skillnad från Diderichsen (1946) har bokstaven S blivit N här: *Jeg foreslår at bruge N (nominalt led eller nominal) i steden for S (substantiviskt led eller substantial) (som i EDG) § 64 f.), da det i et ledstillingsskema er praktisk at slå prædikativer sammen med objekter og de førstnævnte ikke bør kaldes substantiviske* (Diderichsen 1966) (s. 369). Detta kan jämföras med Telemans (1966) uppställning i Tabell 17.

funnit en nivå av abstraktion där enheterna kan beskrivas med hjälp av bara tre strukturtyper och fastställt de ordningsrelationer som gäller för dessa i språket.

Med satsschemats information blir hjälpverbskonstruktioner och därmed icke-finita verb på huvudsatsnivå särskilt eftertraktade. Detta beror på att om positionerna för finit och infinit verb i huvudsatsen har innehåll, avgränsar dessa konstituenter mittfält och fundament på ett tydligt sätt. Med ett avgränsat mittfält gäller oftast den inledande tumregeln i Formel 5 för huvudsatser, i aktiv diates, vilket kan ses som en konsekvens av satsschemat, i modern svenska.

Formel 5 Om ett nominalt led finns i mittfältet i en huvudsats med hjälpverbskonstruktion är detta subjekt
Om det inte finns något sådant där är subjektet i fundamentposition

Att en huvudsats har en hjälpverbsform underlättar således hela analysförloppet. Att kunna markera subjekt med säkerhet är positivt för hela analysen eftersom det undanröjer en nominal kandidat som skulle kunna vara objekt/predikativ. Eftersom metodiken hanterar ett begränsat antal nominala led i varje sats har varje säker uppmärkning dubbelt värde på vägen mot fullständig analys av huvudsatsen. Hjälpverbskonstruktioner är gynnsamma från ett funktionellt analysperspektiv, såvida de verkligen fungerar enligt ovanstående. Undantaget härifrån är de nominalfraser vars huvudord möjliggör att de fungerar som adverbial (oftast tidsadverbial), som i Ex 22 a), som negerande pronomen som i Ex 22 b) och c) och i äldre svenska även objekt som i Ex 22 d).

Ex 22 a) Den förste professorn i ämnet, Jakob Fredrik Neikter (1744-1803), hade ett par år tidigare blivit universitetsbibliotekarie [...] (ja04-050)
b) Men Barbro hade ingenting sagt om att hon skulle åka. (kl15-278)
c) Jag har ingenstans kunnat sätta mig. Börjars o.a. (2003)
d) "Konungen bör rätt och sanning styrka och befordra, vrångvisa och orätt hindra och förbjuda..." (kk01-058)

Subjektet kan även placeras i slutfältet som i Ex 23. Detta kan bero på att övrig kandidat saknas, ofta för passiva satser som i a), men även för vissa aktiva som i b). Exemplet c) är ett exempel på strandad preposition och rektionsframflyttning. (Diskontinuerliga satsled är dock inte en svårighet som är specifik för hjälpverbskonstruktioner.)

Ex 23 a) För genomförande av reformen kommer inom kort att bildas bl.a. en central organisationskommitté. (ha05a-052)
b) Av bekräftelsen skall framgå värdepappersfondens beteckning […] (ha26-208)
c) Att hällkonst utgör ett slags rituella produkter behöver inte råda något tvivel om. (ja26-027)

Definitionsfrågan

Beaktat dessa undantag ses hjälpverbskonstruktioner ändå som särskilt gynnsamma i schemaanalysen. Frågan är i så fall bara vad som ska ses som ett hjälpverb. Teleman (1994) ställer i sitt resonemang upp åtta väl formulerade rent syntaktiska kriterier för prototypiskhet hos svenska hjälpverb (nedan fritt omformulerat), som resulterar i att *skola, lär, torde, böra, måste* och *kunna* blir mest typiska.

Definition 1 Prototypiska hjälpverbsegenskaper enligt Teleman (1994)

Det typiska hjälpverbet kännetecknas av:

1. Oförmåga att konstrueras med *att* (* *Han bör att sjunga*, jfr *Han kommer att sjunga*)
2. Oförmåga att passiviseras (* *Omarbeta det villes inte av dem*)
3. Oförmåga att parafraseras med vanlig (icke-pronominell) propositionell NP (*Han borde – resa till Eslöv/* en resa till Eslöv*, jfr *Han planerade – att besöka sin mor/ett besök hos modern*)
4. Oförmåga att vara korrelat till *göra* (*Han måste vara hemma oftare än jag - * gör / måste*)
5. Förmåga att vid utbrytningsparafras stå i utbrytningens överordnade sats (*Det verkar vara lastbilen som Per tog ↔ Per verkar ha tagit lastbilen,* jfr * *Det försökte vara Per som körde lastbilen =//= Per försökte köra lastbilen*)
6. Förmåga att med oförändrad betydelseeffekt kvarstå om satsen passiviseras (*Per ska måla grinden ↔ Grinden ska målas av Per*, jfr *Per orkade måla grinden =//= Grinden orkade målas av Per*)
7. Förmåga att ta expletivt subjekt (*Det ska komma många* jfr * *Det orkade komma många*)
8. Defekt och/eller oregelbunden böjning (*torde:* defekt böjning, *kunna*: oregelbunden)

Det ska här konstateras att parsningssystem för svenska troligen sällan använder en så sträng definition av hjälpverben. Definitionen är användbar för att inordna ordtypen i ett teoretiskt regelsystem, men för att göra satsschemaanalys kan den möjligen vara väl exkluderande. Frågan som här istället ska diskuteras är: *Hur vagt kan hjälpverben definieras?* Syftet är alltså att så ofta som möjligt skapa ett avgränsat mittfält. Mot ovanstående noggranna uppställning ska här experimentellt presenteras en definition av en gruppering som kraftigt utökar mängden av de ord som ses som finita ’hjälpverbsliknande ordformer’ enligt schemat, och som knappt behöver uppfylla något av de ovanstående kriterierna. I brist på bättre namn kallas den pragmatiskt motiverade gruppen här ’finita hjälpverbsliknande verb’ utan att det på något sätt är syftet att kritisera den normala, eller enligt Definition 1 välformulerade, betydelsen av termen hjälpverb.[33]

[33] Teleman (1994) avslutar sin uppsats med att trots dessa tydliggjorda kännetecknen för den annars oklart avgränsade kategorin låta bli att förkasta namnet hjälpverb, men påtalar att otil-

Definition 2 En experimentell definition av finita *hjälpverbsliknande* verbformer av analysskäl

1. Det finita verbet ska konstruktionsmässigt fungera med ett icke-finit verb så att Formel 5 ovan för subjektsidentifikation blir användbar (med nämnda undantag i Ex 22 och Ex 23, samt för NP-formade adverbial).
2. Verbet ska *relativt ofta* förekomma med verbfraskomplement likt 'normala' hjälpverb[34]

Punkt 2 är ett begränsande kriterium som är satt för att inte helt släppa tyglarna fria: när en stor mängd verb markeras som hjälpverb riskerar icke-primära icke-finit också att ses som primära av misstag. Den radikala definitionen öppnar dörren för en rad nya verb om syftet som nämnts bara är analystekniskt. Från Teleman (1994) accepteras de halvgoda kandidaterna fenomenverben, *verka, förefalla, tyckas, synas* och *befinnas,* och de verb som kallas aspektuella: *börja, råka* och *tendera*. Men dessa 'närhjälpverb' är som ska visas bara början. Det som kan ses som positivt med tolkningen är att analysen enligt satsschemat kan bli betydligt mer detaljerad, enligt Tabell 15. Vad som enligt normal analys är en infinitivfras med objektsfunktion som oavsett längd placeras i N-positionen blir istället en uppdelning i V och andra positioner. Men det innebär också vanskliga förändringar som bör beaktas.

F	v	n	a	V	N	A
a) Han	*beslutade*	*[-]*			*att köpa båten*	
b) Han	*beslutade*	*[-]*		*att köpa*	*båten*	
a) Den	*tålde*	*[-]*	*ju*		*att köras i regn*	
b) Den	*tålde*	*[-]*	*ju*	*att köras*		*i regn*

Tabell 15 Om mindre prototypiska finit som *beslutade* och *tålde* accepteras som hjälpverbsliknande ordformer och det påföljande verbet därför placeras i positionen för icke-finit verb, som i de tonade b)-tolkningarna ändras placeringarna. Det som sker är att 'infinitivfrasens komplement' tolkas som primära led. Detta får ses som en aningen tveksam process. Det kanske ändå ska noteras att dessa 'nya led' oftast kan spetsställas precis som ordinarie primära led.

låtna generaliseringar riskerar att bli följden utan bra definition – vilket är en relevant kommentar i det aktuella sammanhanget.

[34] Det kan nämnas att inte ens de mest typiska ord som kan fungera som hjälpverb verkligen kräver ett icke-finit verbkomplement och hjälpverbsfunktion i svenska: *Vi ska till Skåne, Vi lär det*.

(Van-)tolkningen i Definition 2 innebär att flera klart ovanliga verb som fungerar med infinitivfras som komplement inkluderas som i Ex 24.

Ex 24 a) Nästa år beräknas förlusten ligga på minst 70 miljoner kronor. (aa02b-042)
b) Denna teknik anses framför allt kunna bli betydligt billigare än sedvanlig läkemedelsproduktion. (aa04c-012)

Härmed inbegrips i så fall också verbformer, inklusive passiva, som bildar vad som i t.ex. Ljung och Ohlander (1971) kallas *subjekt med infinitiv (Hon anses vara intelligent)*. När det istället gäller konstruktionstypen *objekt med infinitiv* (*Hon anser henne vara intelligent*) blir konsekvensen däremot att verb(-former) som kan konstrueras så, t.ex. *anser*, inte uppfyller första punkten i Definition 2, och dessa former räknas således inte. Nytolkningen av vilka finita verb som gör att det icke-finita verbet kan ses som placerat i V-positionen i schemat och skapa en hjälpverbskonstruktion får ses som en sorts avsteg. Poängen är dock att göra subjektsidentifikation enkel. Faktum är att det efter att subjektet identifierats är möjligt om så önskas att återgå till den gängse, mindre detaljerade analysen igen (dvs. att placera leden från infinitivkomplementet i N-positionen och benämna dem ’objekt’). Det är också möjligt att plocka bort de extra hjälpverben från programmets gränssnitt. Trots denna utökning av hjälpverb på typnivån är den överväldigande majoriteten av faktiska förekomster av finita hjälpverb i text ändå de prototypiska *har, hade* och *ska/skall*. Den uppsättning av verbformer som användes finns i Appendix.

---

8 979 av **40 000,** dvs. **ca 22,4 %** slumpvis analyserade s-enheter markerades som innehållande *minst ett primärt icke-finit verb (hjälpverbskonstruktion) med den aktuella tolkningen av hjälpverb*.

Frekvensuppskattning 6 Värdena gäller hjälpverbskonstruktioner som markerades med aktuella inställningar och hjälpverb eller hjälpverbsliknande ord enligt Appendix.

7 500 av **40 000,** dvs. **ca 18,8 %** slumpvis analyserade s-enheter markerades som innehållande *minst ett primärt icke-finit verb (hjälpverbskonstruktion) med Mambans hjälpverbsdefinition*.

Frekvensuppskattning 7 Värdena gäller hjälpverbskonstruktioner som markerades med hjälpverbsuppsättningen från Mamban (Teleman 1974), s. 48. Finita former av källans uppställning är de 30: *behöver, behövde, bör, borde, får, fick, kommer, kom, kan, kunde, låter, lät, lär, må, månde, måste, måtte, ska, skall, skulle, torde, tycks, tycktes, tör, verkar, verkade, vill, ville* och dessutom *har, hade.* Verbformerna *få, fick, låter, lät, tycks, tycktes* är speciella eftersom de inte tydligt gör nominalt led i mittfält till subjekt per default. (Även *kom/kommer* har liknande undantag i konstruktioner som *det kom mig att småle*.)

Hjälpverb med particip (*De blev hämtade*) räknas däremot inte in i gruppen hjälpverb alls i den aktuella implementationen. Det beror på att komplementet inte är taggat som verb, vilket tydligt skulle föranleda placering i V-positionen, utan istället som particip, vilket är en egen ordklass i SUC och mer typiskt objekt/predikativ. Att denna konstruktionstyp kan hjälpa till i subjektsidentifikationen är inte desto mindre användbart men det sker med senare specialregler. Något tydligt avgränsat mittfält skapas inte riktigt på samma sätt som med ovan nämnda hjälpverbskonstruktioner.

2.4.2 Metod och resultat

Icke-finita verb (verb i infinitiv eller supinum) är enkla att känna igen genom den morfosyntaktiska taggningen. Dessa är satsled av begränsad längd och de har ett tydligt beroende, nämligen av ett föregående primärt finit verb som är av typen hjälpverb, eller hjälpverbsliknande verbform (enligt den varierande tolkningen). Detta gör att det kan genomföras en process mycket liknande licensiering för att knyta icke-finita verb till primär eller lägre nivå. I likhet med finitet är det icke-finita verbet, som alla led, möjligt att samordna på olika nivåer vilket också för denna uppgift kan erbjuda svårigheter. Svårigheterna som uppkommer genom strykning av *har* och *som*, går igen också för det icke-finita verbet. I övrigt används samma licensieringsprocess för icke-finita verb som för de finita.

I en sökning efter 100 s-enheter från testmängden med minst en markerad primär hjälpverbskonstruktion (Mambans definition användes) räknades 91 som helt riktiga, 4 borde ha licensierats på olika grunder, 3 fel berodde på samordning av primära icke-finita verbfraser och skulle lätt kunna rättas till givet mer tid, ett genererade grafikfel och ett berodde på antaget fel i SUC:s taggning.

2.5 Identifikation av begränsade adverbial, partiklar och reflexiver

Nästa steg i analysen innebär identifikation av övriga led av begränsad längd som utgör fristående led på huvudsatsnivå. Först av dessa tas adverbial upp. Dessa är led som antingen kan spetsställas ensamt med bibehållen satsbetydelse (proposition) som *Jag tror ändå att han vann – Ändå tror jag att han vann*, eller att de när de angränsar till led som kan spetsställas vid denna spetsställning inte följer med: *De vann ju en miljon – En miljon vann de ju*.[35]

Adverbialled över huvud taget har typiskt formen av prepositions- eller adverbfras. En del av dessa, inklusive många mittfältsadverbial, identifieras preliminärt här tidigt och genom matchning. Den andra typen av adverbial inbegriper även objektliknade adverbial (’prepositionsobjekt’) och identifieras enligt den enkla grundregeln *primära fristående prepositions- och adverbfraser är adverbial*. Dessa obegränsade adverbial känns igen senare i processen, i samband med identifikationen av de obegränsade leden i kapitel 3. Det följande stycket tar upp de begränsade adverbialleden, bland vilka (somliga) satsadverb utgör en stor del. Adverbial är i de flesta fall fakultativa (tilläggsled) men i somliga fall egentligen obligatoriska (fyllnadsled: *Han mår utmärkt*). Detta avsnitt gäller identifikationen av de korta adverbial som ofta placeras i mittfält, har begränsad längd (i allmänhet ett ord långa) och i hög grad är satsadverbial, som i Ex 25.

Ex 25 a) Efterfrågan kommer då att överstiga utbudet. (ha14-060)
b) Deras lägenheter blir därför inte lediga i samma utsträckning som förr. (fc03a-012)

Dessa första adverbial som också identifieras genom kandidatinsamling och licensiering är ettordsuttryck eller i vissa fall sammansatta flerordskonstituenter som *t.ex.*[36] Grundvillkoret är att de är ordklasstaggade som adverb (*AB*). Konstituenten som alltså generellt inte har något syntaktiskt beroende av andra led är identifierbart genom att löpordet matchas mot en listning.

I nästa kapitel beskrivs hur analysen av rekursiva led producerar segment (chunkar) som ibland likaledes identifieras som primära adverbial. Exempelvis adverbialet *ännu,* i en egen chunk och med ordklassen adverb, skulle därmed kunna identifieras korrekt i en struktur som *Han ska ännu planera festen* utan att det hanteras speciellt i en listning eftersom det följs av ett annat begränsat led. Po-

[35] En annan, placeringsgrundad, uppdelning av adverbialen skisseras i Diderichsen (1966) och sker på grundval av möjlig placering. Här finns uppdelning bland typiska adverbial i mittfältsposition osv och de olika begränsningar för flyttning till olika positioner som gäller (Se Kapitel 5). Den metod för identifikation som här beskrivs undersöker alla satsfält.

[36] Att *”t.ex.”* grafiskt måste skrivas som precis ett ord utan mellanslag är diskutabelt – inte desto mindre förekommer grafordet så i SUC 2.0 och har där bl.a. denna roll.

ängen med att identifiera dessa adverbial i detta tidiga skede är dock att i andra fall avgränsa segment i satsen, t.ex. i ett fall som *Han planerar ännu festen*. Adverbialidentifikation utvärderas i kapitel 3.

När det gäller reflexivet *sig* och partiklar så följer metoden verblicensieringens procedur på ett okomplicerat sätt. Kandidater insamlas och licensiering genomförs på samma vis. Partiklar känns igen genom taggningen (*PL*) medan reflexiva pronomen bara speciellt identifieras när de har formen av ordet *sig*. Andra former av reflexiv känns inte igen på detta speciella sätt. Dessutom ska tolkningen som reflexiv uteslutas i konstruktioner som *vare sig, i och för sig* etc. Den praktiska skillnaden som denna feltyp leder till verkar främst vara att reflexivet inte kan spetsställas. Idealiskt, dock ej helt utarbetat här, skulle valensinformation hos verbet utesluta endera tolkningen.

Verbpartiklar (PL), dvs. betonade partiklar, har i vissa versioner av schemat en speciellt utsatt position. För identifikationen av primära partiklar, enligt den aktuella ansatsen, har positionen dock ingen betydelse, förutom när det gäller allmänna begränsningar som att det ej finns ensamt spetsställt. Identifierade verbpartiklar är användbart i den senare subjektsidentifikationen, men på varierande sätt beroende på verb och partikel: exempelvis *NOM ser NOM ut* respektive *NOM kommer NOM på* innebär per default olika tolkningar angående var subjektet finns.

Resultaten för identifikation av de primära av dessa begränsade led är i princip helt utbytbar mot frågan om programmet är kapabelt att känna igen kandidater och avgöra ifall en sekvens finns på primär eller lägre nivå. Med andra ord är proceduren för adverbial, och även för partiklar och reflexiver, likadan som för verben: det sker en insamling av kandidater och därefter licensiering.

---

5 467 av **40 000,** dvs. **ca 13,7 %** slumpvis analyserade s-enheter innehöll minst en verbpartikel.
2 965 av **40 000,** dvs. **ca 7,4 %** slumpvis analyserade s-enheter var markerade av syntaxanalysatorn som innehållande minst en primär verbpartikel.

Frekvensuppskattning 8 Förekomsten av (betonade) verbpartiklar på primär nivå.

I en sökning efter 100 framslumpade s-enheter med minst en markerad primär verbpartikel från testmängden bedömdes tre vara fel (icke-primära) och berodde

på missad *som*-strykning. En s-enhet bland de som sågs som korrekta borde ej vara märkt som partikel alls i SUC.[37]

---

2 536 av **40 000,** dvs. **ca 6,3 %** slumpvis analyserade s-enheter innehöll minst en förekomst av *sig*.
1 383 av **40 000,** dvs. **ca 3,5 %** slumpvis analyserade s-enheter var markerade av syntaxanalysatorn som innehållande minst en primär förekomst av *sig*.

Frekvensuppskattning 9 Förekomsten av *sig* generellt och på primär nivå.

I en sökning efter 100 framslumpade s-enheter ur testmängden med markerat primärt *sig* bedömdes två som fel och borde licensierats med en *som*-strykning som ej identifierats.

2.6 Identifikation av primära konjunktioner

Som *primära konjunktioner* räknas de konjunktioner, kommatecken, tankstreck, kolon etc. som samordnar primära satser och primära verbfraser. Den första gruppen är sådana som samordnar huvudsatser (Ex 26 a och b) och primära *finita* verbfraser (Ex 26 c och d). Grundregeln är som nämnts att varje primärt finit verb ska vara ensamt i sin huvudsats/primära finita verbfras. Mellan dessa två konjunkttyper finns bara skillnaden att huvudsatser generellt har fundament och oftast ett subjekt. Det är därmed placeringen av den primära konjunktionen och den uppdelning som sker som avgör om konjunkten är huvudsats eller finit verbfras. Den efterföljande identifikationen av de obegränsade leden (se Kapitel 3) sker i det område som avgränsas av de primära konjunkterna, och förfält när sådana finns.

Ex 26
a) Det är sällan några minnesvärda matcher så jag kommer inte ihåg dem men jag stirrar i alla fall. (kr01c-006)
b) Den startade sannolikt ca 1905-1910 men modellernas och konstruktionernas inbördes kronologiska ordning är långt ifrån utredd. (ea21-095)
c) Den vita katten med violblå ögon tryckte sig mot dörren och hoppades förgäves på att bli insläppt i stugvärmen. (kk03-092)
d) Han granskar förhörsprotokollen och ställer sig överhuvudtaget kritisk till hela rättegångsförfarandet. (ga06-087)

[37] ec06-012: Bolaget skulle därför sträva *efter* att […].

6 602 av **40 000,** dvs. **ca 16,5 %** slumpvis analyserade s-enheter var markerade av syntaxanalysatorn som innehållande *minst en primär konjunktion*.

Frekvensuppskattning 10 Förekomsten av primära konjunktioner beräknas till 16,5 %. Jämför Diagram 7 om frekvens för s-enheter med olika antal primära finita verb.

Identikation av primära konjunktioner görs genom att tilldela konjunktioner mellan två primära finit olika poäng beroende på kontext och position, även själva ordet, t.ex satssamordnaren *ty,* som särskiljs bidrar till poängen. Identifikationssäkerheten för primära konjunktioner testades genom att söka efter 100 s-enheter, i testmängden ur SUC, med minst en utsatt primär konjunktion. Av 100 enheter som markerats innehålla mer än en huvudsats/primär finit VP var en klar majoritet av valen enkla p.g.a. att bara en kandidat för primär konjunktion fanns. Fem enheter antogs vara felanalyserade så att t.ex. en finitlicensiering (t.ex. *som*-strykning) missats och det blev därmed inte fråga om en här eftersökt samordning. Fyra enheter antogs blivit fel p.g.a. fel taggning i SUC, vilket bl.a. inneburit att egentlig licensierare (t.ex. *som, där*) hade en taggning som icke-bisatsinledare. I 17 av s-enheterna fanns det intressanta läget att mer än en tydlig kandidat till primärkonjunktion förekom, och i 14 av dessa valdes den rätta samordnaren av systemet vid detta tillfälle.

Ex 27
a) Även om motståndarlaget är starkt skall vi kunna behålla vårt eget system och inte behöva virra till det med panikåtgärder. (eb04a-062)
b) Formspråket har på det senaste året reducerats än mer och stramats upp. (cc03d-027)
c) Hela förra vintern hade han arbetat tillsammans med Alvar Yxberg och kört hästforor mellan Hudiksvall och Norge. (kn01-118)

En annan konjunktionstyp som identifieras är den som samordnar *primära icke-finita* verbfraser, enligt Ex 27. Denna primära konjunktionstyp är enkel att identifiera korrekt. Metoden innebär helt enkelt att välja den konjunktion som närmast föregår. Det har varit svårt att hitta undantag här, även om ett adverbial som innehåller konjunktion som i *Han ska skotta och till och med grusa* exemplifierar denna möjlighet.

2.7 Identifikation av förfält

Före fundamentpositionen finns plats för en inledande del som saknar direkt funktionell roll i den påföljande satsen, även om den inte kan kallas betydelselös och i Ex 29 nedan introducerar ämne. Denna sekvens är oftast en konjunktion men den har i vissa fall ett betydligt mer svårfångat mönster. Dessa segment är egentligen inte begränsade i längd men *avgränsande* i likhet med de primära led

som beskrivs i detta kapitel. Eftersom denna ansats har praktiska syften tolkas parenteser och talstreck delvis också som förfältsinnehåll, se vidare nedan.

Ex 28 a) Och därför är, vilket världen känner till, vår litterära kultur världens främsta. (ja21-069)
b) Men det är faktiskt en chimär. (gb18-082)
c) (Detta gäller vid låga energier, c:a 1 MeV. (fh10-074)

Ex 28 exemplifierar de vanligaste typerna av förfältsinnehåll. De är frekventa och relativt lätta att identifiera genom matchning i områdena direkt i textmeningens start eller i området omedelbart före det primära finita verbet. Fallen i Ex 29 tas istället företrädesvis hand om genom en matchning av själva fundamentledet. Förfältsinnehåll går under namn som *initialt annex* i SAG (Teleman, Hellberg och Andersson 1999) respektive *fria fundament* eller *initialt extraponerade led* i Andersson (1994).

Utvidgad sats				
Förfält	Inre sats			Efterfält
	Initialfält (Fundament)	Mittfält	Slutfält	
Imorgon,	*då*	*kan hon nog*	*vara med,*	*din syster.*

Tabell 16 Beskrivningen av den utökade satsen enligt SAG (4:6) inbegriper fältpositioner för initiala och postponerarade annex i efterfält. Huvudfokus för arbetet är 'den inre satsen'. Den utökade satsbeskrivningen används kanske framförallt i talspråksanalys.

Ex 29 a) "Hejsan Nisse, jag skulle behöva låna en halv miljon." (kr05-071)
b) Hon med sin rikssvenska sörmländska och han med sin nordskånska skorrning - jag minns knappt att de talade med varann. (fc05-135)
c) - Mycket intressant, herr Selander, men den här turen tar lång tid. (kn06-106)

Ex 29 a) och b) illustrerar att 'avgränsare – personligt pronomen – primärt finit verb' i en del lägen skiljer huvudsatsens fundament från en tidigare sekvens (här motsvarar denna sekvens den kommande bisatsens subjekt, men räknas inte som del av satsen syntaktiskt). Det ovanligare fallet i Ex 29 c) verka kräva en sorts igenkänning av den bestämda nominalfrasen *den här turen* för att kunna utesluta föregående segment.

3 525 av **40 000,** dvs. ca **8,8 %** slumpvis analyserade s-enheter var markerade av syntaxanalysatorn som innehållande *förfältsmarkerade ord/tecken.*

Frekvensuppskattning 11 Förfältsidentifikation som avgränsar framåt i s-enheter markeras i 8,8 % av s-enheterna.

Förfält identifieras dels genom matchning, så att vissa inledande konjunktioner och vissa sekvenser inte hamnar i fundament eller på primärfinits position. Dessutom används en s.k. negativ definition där istället själva fundamentet matchas och den föregående resten, utan djupare analys, tolkas som placerat i förfältet. Ex 29 a) visar ett fall som analyseras korrekt genom att pronomen med subjektskasus mellan kommatecken och primärfinit tolkas som ensamt subjekt. Programmet identifierar generellt talstreck och parentesstart som förfältsinnehåll, däremot inte citattecken eller talstreck (eller liknande streck) som föregår yttrande i s-enhet med direkt anföring med angiven yttrare - *Det var bra, sa han*.

Hur väl fungerar den beskrivna metoden, och hur ser fördelningen av faktiska förfältssegment ut i svensk publicerad text? En mindre undersökning genomfördes också för att ge en bild av korrektheten (närmare bestämt precisionen) av förfältsidentifikationen samt hur de olika förfältsinnehållen fördelades formmässigt. En sökning gjordes efter 100 framslumpade s-enheter med förfältsinnehåll – uppenbart kan av programmet inte identifierade förfält ha missats. Erfarenhetsmässigt är de som missats att identifieras långa led, eller kombinationer av typiska förfältare. Av de 100 markerade enheterna betraktades tre som felanalyserade förfält. De övriga fördelade sig enligt Diagram 8.

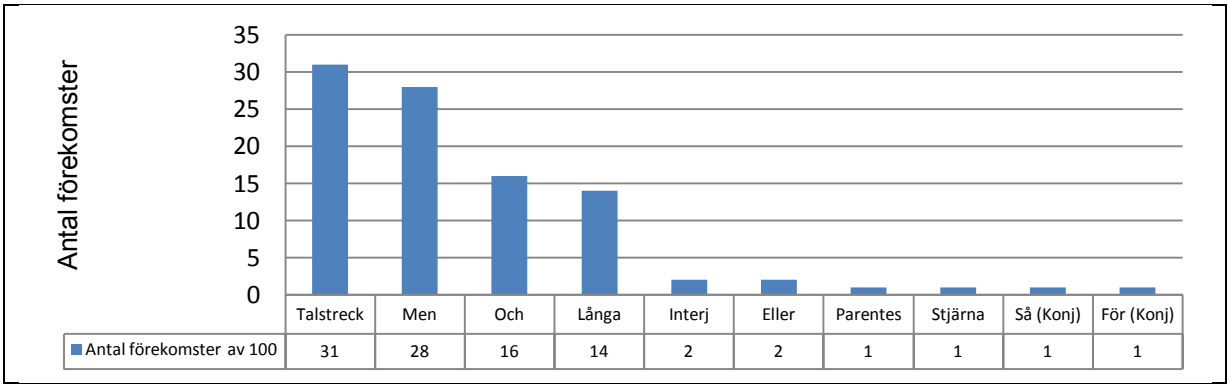


Diagram 8 Av de 97 av 100 korrekta uppmärkningarna utgjordes majoriteten av ettords- eller etteckenuttryck. Konjunktionen *Men* (ensamt) är det vanligaste ordet som fungerar som förfältare. Kategorin *Långa* innehåller kombinationer av uttryck som *interjektion – komma* och fria annex som *Stundtals faktiskt hejdlöst rolig: aldrig har väl* […]. Talstreck innebär inte nödvändigtvis ett typografiskt långt eller speciellt markerat tecken. 33 av de markerade s-enheterna här gäller teckenformade segment i stället för ord.

Sammanfattningsvis kan sägas om identifikationen av de primära begränsade satsleden i satsschemat att kandidater till dessa generellt kan samlas in enkelt genom ord- och taggningsinformation medan licensieringsprocesserna inte riktigt hundraprocentigt identifierar varje förekomst som primär eller syntaktiskt underordnad.

3 Identifikation av obegränsade primära satsled

Efter identifikation av begränsade primära led har textmeningar delats upp i huvudsatser samt primära finita verbfraser avgränsade med primära konjunktioner. Dessa segment är i sin tur uppstyckade genom identifikationen av primära verb, verbpartiklar, reflexivpronomen, vissa adverbial, förfält och samordnare av infinita verbfraser. Detta utgångsläge möjliggör en analys som i många fall innebär att bara ett led av obegränsad längd finns i vart och ett av de kvarvarande områdena, särskilt i fundamentet. En tidig version av delar av innehållet i detta kapitel, framförallt *rangbaserad chunkning*, presenterades i Wilhelmsson (2008).

Ex 30
a) *Han* **bad** *om ursäkt* **och gick och la sig.** (kl05-209)
b) *Garnisonen på Karlavägen i Stockholm* **revs** *efter det att SCB lades ned***.** (hb01b-098)
c) **-** *Var* **ligger** *det***?** (kn16-096)
d) *I de nyupptäckta fullerenernas värld* **är** *det en realitet***.** (fh04a-003)

Ingångsläget för den uppgift som beskrivs i detta kapitel är alltså den uppdelning i satser, verbfraser och fält som förra kapitlets identifiering av huvudsatsnivåns begränsade led skapat. Ex 30 exemplifierar hur detta utgångsläge ter sig i några olika konstruktioner där de begränsade primära leden (i fetstil) delat upp huvudsatser, primära verbfraser och fälten (*fundament-, mitt- och slutfält* samt *efterdel*, dvs. delen efter finit i icke hjälpverbskonstruktion). Detta kapitel beskriver uppdelning och identifikation av de obegränsade funktionella leden (understrukna i exemplet), dvs. de som saknar övre längdgräns (subjekt, objekt/predikativ och adverbial). I den aktuella metoden innebär denna uppgift ofta en mycket annorlunda uppgift än hur den framstår jämfört med en parsningsalgoritm som använder t.ex. en kontextfri grammatik. Genom tumregler som att fundamentet i princip bara innehåller ett led, är det ofta fråga om enkla uteslutningsregler i den programmerade metoden. För att avgöra gränser mellan angränsande rekursiva konstituenter i samma fält är det dock nödvändigt med en sorts segmentidentifikation, här i form av en presenterad metod som kallas *rangbaserad chunkning*. Vidare används för avgörandet om prepositionsfraser är attribut- eller adverbialled data från av två stora valenslexikon. I detta steg av analysen finns också många heuristiska regler som främst sammanfogar ’chunkar’ till större segment på grundval av bl.a. angränsande ord i dessa och en stor samling manuellt skapade regler. En tidig grov uppskattning av konstruktionstypers vanlighet, med metodens angreppssätt i åtanke, återfinns i Diagram 9.

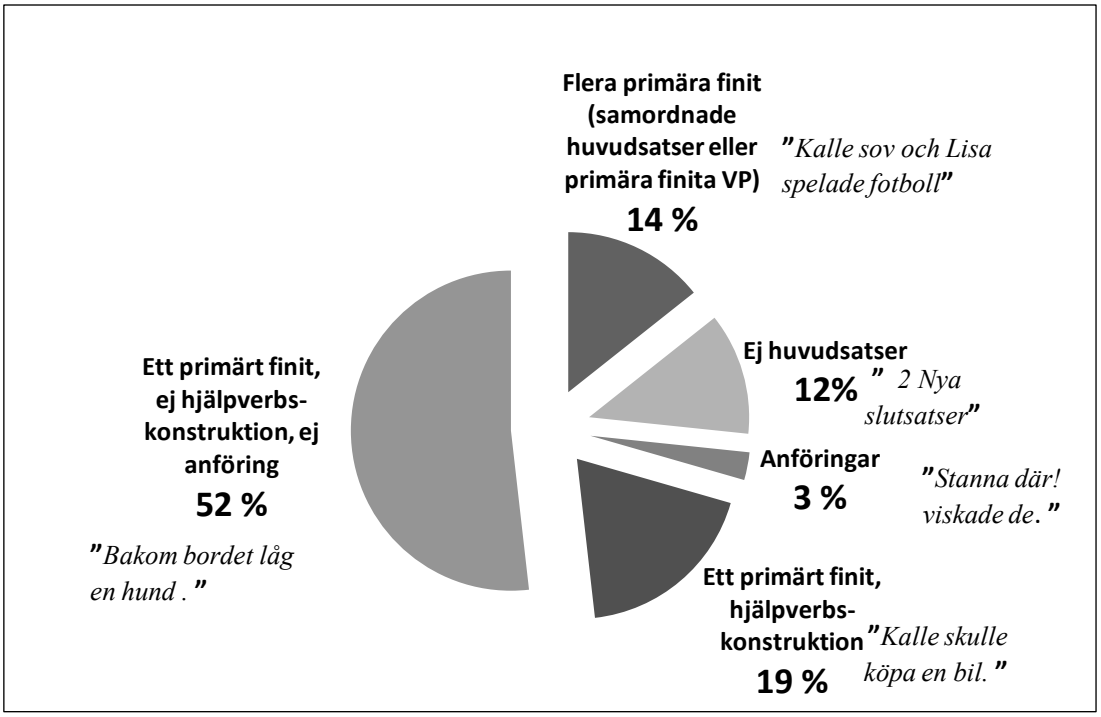


Diagram 9 Uppskattningen av konstruktionstypers relativa andelar utgår från en tidig manuell undersökning av 805 slumpvis utvalda SUC-enheter. Enligt denna uppskattning är ungefär hälften av enheterna i SUC enkla huvudsatser (med ett primärt finit), ca 19 % är primära hjälpverbskonstruktioner med bara ett primärt finit, ca 14 % innehåller mer än ett primärt finit, 12 % är icke-huvudsatser som rubriker, fristående bisatser etc. som räknas som s-enheter och 3 % är anföringar.

Struktur i kapitel 3, Identifikation av obegränsade primära satsled

I avsnitt 3.1 beskrivs hur de segment som identifieras i de avgränsade fälten kan besätta rollerna som subjekt, objekt/predikativ och adverbial utgående från rent strukturmässiga kriterier. Avsnittet beskriver det allmänna förhållandet mellan syntaktisk strukturtyp (huvudsakligen *frastyp*) och funktionell kategori.

Avsnitt 3.2 tar upp den inledande formationen av chunkstrukturer i de fält som avgränsats av de begränsade leden och förfält. För detta ändamål introduceras en ny chunkningsmetod, *rangbaserad chunkning*, som använder förhållandevis lite information i den ordklasstaggning med särdragsvärden för varje löpord som SUC 2.0 tillhandahåller.

Avsnitt 3.3 visar en jämförelse mellan den presenterade metoden för chunkning och några system för svenska med liknande syften.

Avsnitt 3.4 beskriver hur de bildade segmenten (chunkarna) stegvis sammanfogas för att möta ett antal nominala led (subjekt och objekt/predikativ) som är kompatibelt med satsform och verbvalens. Den stegvisa sammanfogningen sker

genom manuella regler, som ofta sammanfogar på grundval av angränsande ord i två angränsande segment. Dessutom används valenser för substantiv, adjektiv och particip från valenslexikon och andra heuristiska regler.

Avsnitt 3.5 beskriver etiketteringen av de bildade segmenten med funktionella kategorier: subjekt, objekt/predikativ och adverbial. Olika satsstrukturer hanteras med skilda strategier då exempelvis hjälpverbsstrukturer bjuder på klart speciella förutsättningar. Temat *identifikation genom uteslutning* märks här genom att genomförd subjektsidentifikation generellt lämnar kvar andra nominala strukturer som kandidater till primära objekt/predikativ.

Avsnitt 3.6 tar upp ett antal speciella svenska satskonstruktioner som av olika skäl är svåra att täcka eller kräver speciella regler. En eventuell rigid tolkning av satsschemats innebörd gör dessutom att vissa konstruktioner inte riktigt passar in. Dessa eventuella undantag är oftast mindre frekventa och välkända.

Avsnitt 3.7 innehåller ett försök att relatera resultaten till liknande uppgifter i andra parsningssystem för svenska.

3.1 Förhållandet mellan struktur och funktionell kategori

I följande avsnitt kommer redogöras för hur frasstrukturer som motsvarar de primära obegränsade leden subjekt, objekt/predikativ och adverbial identifieras genom *rangbaserad chunkning* och ytterligare sammanfogning av attribut till större frasenheter på grundval av valensdata och heuristiska regler.

Den första frågan att reda ut gäller vilka strukturer som uppbär rollerna subjekt och objekt (samt eventuellt predikativ). Denna besvarades på ett klargörande sätt och utgående från satsschemat av Teleman (1966), s. 113–115.[38] Det bör dock betonas att denna uppställning efter distribution gäller subjekt, objekt och nästan alltid prepositionskomplement (rektion) men att strukturer som utgör predikativ utesluts från gruppen. Den gruppering som Diderichsen gör mellan adverbiella och nominala led innebär ju att objekt och predikativ hamnar i samma grupp och kallas ’N’, se Formel 4. Kopulaverbkonstruktioner särskiljs inte från andra i utdatakodningen.

Förutom kategori E) i Tabell 17, *naken sats*, har strukturerna ovan visat sig svara väl mot de kandidater till de nominala leden som här beaktas, exklusive pre-

[38] Det är också från denna uppsats som termen *nominal* (substantiv: ett nominal) börjar bli vedertaget för svenska. Teleman väljer denna term efter att också diskuterat de möjliga namnen *konstruktion, fras, syntagm, uttryck, helhet, konstitut, konstrukt* och *ordgrupp*.

dikativ. Undantagsvis förekommer också andra led, som utländska uttryck och titlar med alla möjliga strukturer. För att göra bilden komplett ska även nämnas att utsagan i ett rapportuttryck alltså räknas som objekt, men den känns inte igen genom struktur.

Struktur	Exempel
A) Substantiv eller konstruktion med substantiv som kärna	*Alla mina många andra sådana nya danska förtjusande böcker; general Almgren; en riksdagsman och en redaktionschef; antalet studenter; nummer 26; klockan 3.00; svensk tid; en generell höjning på 8-10 procent; vägen dit; jobbet som biträde; far min; ett förhållande varom man idag näppeligen kan få några upplysningar; mat att äta*
B) Liknande (NP-)struktur utan substantivkärna	*Alla dessa; många andra; det mirakulösa med hans konst; en av landets främsta kärnfysiker*
C) Pronomen som står självständigt	*Sig själv; vi alla; allting annat trevligt; jag ensam*
D) Infinitivfras	*Att nu tillämpa en sådan bedömning; att emigrera*
E) Naken sats (en sorts att-strykningsrest)	*[du hade bett] vi skulle komma; [det var bra] du kom*
F) *Att*-sats	*Att en sådan expansion inte är tänkbar*
G) Bisats inled med vad som enligt SUC:s taggning är frågande/relativ(t) pronomen/adverb/determinerare	*Hur långt man hunnit i Kina; vad författaren egentligen menat; vilken du vill*
H) *Om*-sats	*Om det sagda äger tillämplighet*

Tabell 17 Uppställningen av nominala strukturer i Teleman (1966), här i lätt modifierad version, är en vägledande grundläggande översikt.[39]

När obegränsade konstituenter på former som NP, PP, infinitivfraser m.fl. identifierats med sin fulla sträckning i de luckor som återstår råder ett visst förhållande mellan strukturtyp och funktionell kategori – enligt nedanstående resonemang. Inledningsvis antogs nedanstående beskrivning över förhållandet struktur – funktion. Att som i Tabell 18 försöka att se prepositionsobjekten som objekt

[39] Om *alla* strukturer som kan utgöra subjekt ska redovisas uttömmande borde troligen många andra bisatstyper, förutom om-satser, också nämnas här, t.ex. *när*-satser som i *När det ska ske är nu klargjort*.

kan tyckas vara märkligt, och är resultatet av en semantisk snarare än syntaktisk kategorisering, på grundval av bl.a. just namnet *prepositionsobjekt*.

	Subjekt/objekt/predikativ	Adverbial
NP, infinitivfraser, att-satser, adjektivfraser	*Normalfallet*	*NP-formade adverbial (oftast tidsadverbial):* • denna gång • kl 16.00 • sommarkvällar • en bit
Prepositionsfraser	*Prepositionsobjekt* Lyssna på en låt av henne Informera om något	*Normalfallet*

Tabell 18 Sammanställningen visar att ett utgångsläge med korrekt identifierade frasstrukturer i den följande analysen ändå, teoretiskt sett, innebär många möjliga slutliga analyser, om prepositionsobjekt, som namnet antyder, tolkas som objekt. (Metod för att finna dessa frasstrukturer som motsvarar primära satsled redogörs för i kommande avsnitt.)

Beskrivningen i Tabell 18 var under inledande analyser de aspekter som ansågs nödvändiga att överväga i identifikationen. Den terminologiska förändringen i svensk grammatikbeskrivning, som återges nedan, förenklar läget.

	Subjekt/objekt/predikativ	Adverbial
NP, infinitivfraser, att-satser, adjektivfraser	*Normalfallet*	*NP-formade adverbial (oftast tidsadverbial, se ovan)*
Prepositionsfraser	*(Inga fall)*	*Normalfallet – Adverbial (inklusive objektsliknande adverbial)*

Tabell 19 Tabellen redogör för det struktur–funktionsförhållande som gäller då prepositionsobjekt kategoriseras som adverbial, i enlighet med t.ex. SAG.[40]

[40] Det är som nämnts svårt att uttala sig kategoriskt med sådana här enkla uppdelningar. Ett undantag från regeln att PP ej är nominal finns i t.ex. titlar, *Till Paris var annorlunda tonsatt*.

Den senare tabellen möjliggör det enkla förhållandet att *fristående prepositionsfraser på huvudsatsnivå är adverbial*. Även om somliga av dessa alltså har ett förhållande som fyllnadsled till verbet innebär denna uppdelning pedagogiska fördelar för analysen: när subjekt och objekt/predikativ ska bestämmas utgör bara övriga nominala strukturer kandidater. Frågan som här ska tas upp är hur 'prepositionsobjekt' som kategori egentligen har hanterats tidigare. Prepositionsobjekten (i Mamban kallat *objektadverbial*) har växelvis sorterats som adverbial och objekt i grammatiklitteraturen under en långvarig diskussion. Anledningen till uppmärksamheten här hänger på ett sätt samman med uteslutningsmetoden: Skulle klassen ses som objekt, som förekommer i begränsat antal per sats, eller som adverbial, vilka är en gruppering med okänt antal förekomster per sats. Det nedanstående är en redogörelse för olika ståndpunkter angående prepositionsobjekts klassificering i den svenska grammatiklitteraturen.

- Hos Thorell (1973) används termen *prepositionsverb* för konstruktionstypen och betyder då verb + preposition (*längta efter*). Den i prepositionsfrasen inledande prepositionen kallas dock *trycksvag verbpartikel* (567). Ett exempel som *Hon drömde <u>om sin ledighet</u>* hamnar mittemellan verb med adverbiell bestämning, som *Han skriver <u>med kulspetspenna</u>,* och verb med tryckstark verbpartikel. I somliga lägen sägs samhörigheten mellan verbet och prepositionen ('partikeln') belysas genom omskrivning till prefixavlett verb: *Hon tvivlar på det ↔ Hon betvivlar det* (684).
- Hos Holm och Larsson (1980) betraktas dessa "mellanting mellan objekt och adverbial" som prepositionsobjekt eller objektadverbial (som i Mamban). Ställningstagande för antingen objekt och adverbial avgörs dock med följade regel: *Om hela den prepositionsinledda satsdel som står efter PRED 2 [dvs. infinit verb i normalposition] inte kan ersättas med adverb och om verbet bestämmer valet av preposition, betraktas satsdelen som OBJ [dvs. objekt]*. Prepositionen som inleder kallas även här *trycksvag verbpartikel*.
- Hos Jörgensen och Svensson (1986) finns prepositionsobjekt kortfattat beskrivet som *den prepositionsstyrda motsvarigheten till direkt objekt* tillsammans med exempel som *Stina ser på Pelle* (s 96).
- Hos Andersson (1994) används termen prepositionsobjekt med invändningen "sådana kunde också räknas som adverbial, eftersom de liknar adverbial till formen och ibland har friare placering efter andra adverbial. Däremot liknar de objekt genom att de som de flesta objekt kan anger föremålet för verbets handling och numera ofta kan bli subjekt i passiv sats." Ett exempel på detta från boken är *Jag kan prenumerera för ett halvt år <u>på Hufvudstadsbladet</u>. ↔ <u>Hufvudstadsbladet</u> kan prenumereras <u>på</u> för ett halvt år.*
- Hos Josefsson (2001) är beskrivningen att prepositionsobjekt är en prepositionsfras som semantiskt fungerar ungefär som ett direkt eller indirekt objekt, dvs. det uppbär samma semantiska roll som sådana objekt. Prepositionsobjekt sägs kunna kännas igen bl.a. genom att rektionsframflyttning (även om

detta inte är helt särskiljande). I denna källa nämns även termen *objektadverbial* för samma led.
- I SAG och i Svenska Akademiens språklära (2003) används termen *objektliknande adverbial* och definitionen sker genom valensbegreppet: *dessa är bundna till verbet genom dettas valens* (s. 238). Att de genom sin struktur ändå räknas som adverbial har varit ett omtalat ställningstagande. Om SAG genom sin dignitet kommer att slå ut begreppet prepositionsobjekt från framtida grammatikbeskrivningar återstår att se.
- *Termlexikon i språkvetenskap* (Åström 2007) har lämpligt nog med båda termerna men låter *prepositionsobjekt* vara en alternativ term för *objektliknande adverbial*.

SAG (Teleman, Hellberg och Andersson 1999) använder alltså inte termen *prepositionsobjekt* utan bl.a. *objektsliknande* adverbial (Band 1, s 203). Ett ställningstagande som tas upp och argumenteras väl för av Hellberg (2003). Att klassificera den omnämnda konstituenten som ett adverbial stöds genom att den kan placeras efter något annat adverbial i efterfältet (slutfältet), till skillnad från objekt. De är bl.a. mindre benägna att finnas i mittfältet.

I de verbvalenser som finns i Nationalencyklopedins ordbok, NEO (1995–96), se kapitel 4, utgör möjliga prepositioner och partiklar en betydande del av informationen – om denna information tas bort består valensen inte av så mycket mer än mono-, di- eller intransitivitet i fråga om nominala led, samt möjliga (tryckstarka) partiklar (dvs. det som i SUC räknas som ordklassen partikel). Oavsett om de prepositionsfraser som hör tydligt samman med verbet på det sätt som här beskrivs kallas för prepositionsobjekt, objektsliknande adverbial eller något annat så görs alltså i de flesta grammatiker en tydlig koppling mellan verb och denna konstituenttyp som särskiljer dessa prepositionsfraser från andra adverbiella bestämningar.

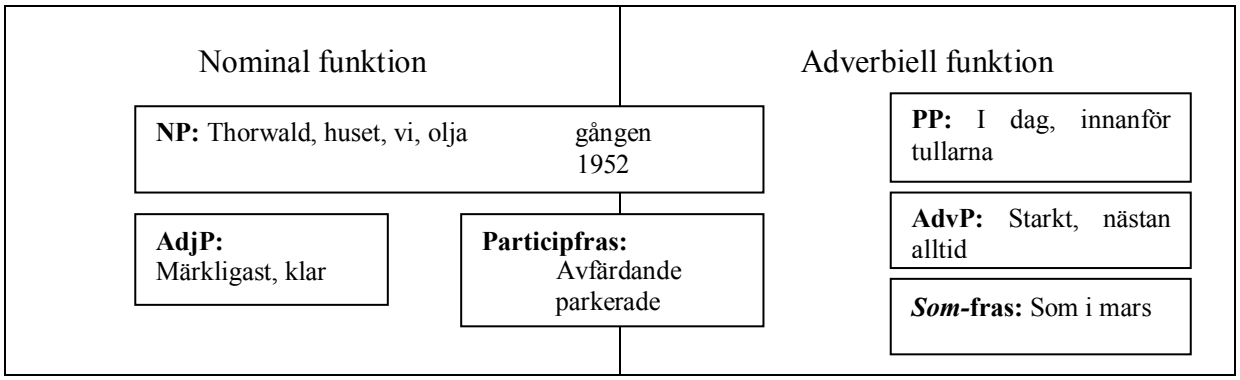
En poäng från litteraturen är att de nämnda prepositionsfraserna är ’objekt’ (när detta är vad de kallas) på *semantiska*, inte rent form- eller placeringsmässiga, grunder och att denna semantik inte verkar direkt åtkomlig i en automatisk analysprocess och att verbet så att säga bestämmer vilka prepositioner som är möjliga att inleda detta led med – strukturen hos själva prepositionsfrasen inklusive den inledande prepositionen i sig ger få ledtrådar om vilken roll prepositionsfrasen innehar. Om prepositionsobjekt/objektsliknande adverbial ska kunna särskiljas från andra adverbial leder dessa förhållanden därför rimligen till att *valensinformation* på något sätt är helt nödvändig från ett datoranalysperspektiv.

Om prepositionsobjekt genomgående, som här, kategoriseras som adverbial och om dessa placeras bland satsschemats övriga adverbial så blir uppgiften att placera olika led på rätt position simplare eftersom objektkategorin i satsschemat

därmed inte skulle kunna upptas av prepositionsfras, vilket möjligen vore en tolkning annars. I valenslexikonen NEO (1995–96) och *Lexin – Svenska ord* (1998) är just prepositionsinformation den största delen av innehållet. Om prepositionsobjekt tolkas och placeras i princip som övriga adverbial så har det ingen betydelse i ett datoranalysperspektiv ifall de är regelmässigt bundna till verbet eller ej, så länge de placerings- och funktionsmässigt ändå inte skulle skilja sig så mycket från övriga adverbial. Den enda kategori som en fristående prepositionsfras på primär nivå skulle kunna inleda blir adverbial.

I den inledande fasen av detta projekt sågs de objektliknande adverbialen som på ett sätt jämställda med andra objekt. Från ett analysperspektiv har det emellertid visat sig enklare att åtminstone inledningsvis behandla dem som andra adverbial, med vilka de allt som oftast delar struktur, som PP. Metoden får därmed en uppdelning som innebär att nominala strukturer inledningsvis eftersökts som kandidater till subjekt och objekt/predikativ medan PP kan undantas.

Fristående NP uppbär oftast rollen som subjekt eller objekt/predikativ. Somliga nominalfraser fungerar potentiellt som adverbial. Det gäller huvudsakligen de med s.k. tidsnomen som huvudord, som *gång, tid, onsdag, 14.50, februari* (uttryck som tenderar att kunna förkortas i skrift på mycket varierande sätt) eller årtal. Andra NP-adverbial inkluderar sträckor som *kilometer* eller *bit* (*Jag åkte bil en bit*) och *grand/grann* (när dessa ord ses som substantiv i SUC 2.0). Dessutom blir *tack vare*-konstruktioner adverbiella, funktionellt sett (fast *tack* i SUC 2.0 oftast taggas som substantiv, och mer sällan som interjektion). Uppställningen av ord som markerar att nominalfrasen kan fungera som adverbial innehåller i skrivande stund ca 265 uttryck (här räknas ordformer för sig). Denna lista med tidsnomen och andra nomen som kan ge frasen roll som andra adverbial finns i Appendix. Syftet med insamlandet har delvis varit att åtminstone *exemplifiera* olika fall av vissa ordtyper, för att möjliggöra t.ex. analys där sådana ord är efterled i sammansättningar, istället för att ge en uttömmande listning. Detta är naturligtvis fallet för årtal och klockslag som rimligen inte listas uttryckligen.



Figur 7 En översikt av tendenser för fristående frastypers roller i ett funktionellt grammatiskt perspektiv innehåller flera exempel på kategorier och enskilda förekomster med möjlighet till både nominal (alltså potentiellt uppbärande subjekts- eller objekt/predikativroll) och adverbiell roll.

Det förhållande som redogörs i Figur 7 är ett försök till generell beskrivning av förhållandet mellan struktur och funktion. Givet den ordklasstaggning som förekommer i SUC och komplicerande fenomen som citat och titlar finns undantag. (Se vidare i 3.7.) Som nämnts är de strukturer som har dubbla möjligheter främst NP med potentiell adverbialfunktion, men även particip. I praktiken finns en stor inverkan från olika konstruktionstyper. Participfraser och adjektivfraser är ofta adverbiella, men i kopulaverbkonstruktioner ofta predikativ. En undergrupp av dem står dessutom ofta för mänskliga referenter (t.ex. *berörda*) och kan lika gärna vara subjekt.

PP-formade subjekt/objekt/predikativ ändå?

Ex 31 a) *I princip* alla näringsidkare som driver tillfällig försäljning skall vid försäljningen vara skyldiga att upplysa om sitt namn, sin postadress och sitt telefonnummer. (ha29-082)
b) - *Omkring* en tredjedel av de förtroendevalda är bussförare. (ec16a-047)

Det finns faktiskt segment som blir märkta som prepositionsfraser med den beskrivna metoden, men som trots det fungerar som nominala led. Detta beror på att fraser som i Ex 31 har ett inledningsord som är uppmärkt som preposition. Det är frågan om ett PP-format framförställt attribut. Som kan noteras föregår prepositionen i exemplet här räkneord eller kvantifikator, och kan i sin tur utgöra komplement till en egentlig PP: *Till omkring en tredjedel* etc. Konsekvensen är att korpusen och därpå grundade ordklasstaggare ger en PP-struktur åt NP i fall när dessa inleds av *omkring, kring, till synes, bland annat, till exempel*.

Eftersom relationerna mellan struktur och funktion som beskrivits här, nämnda undantag till trots, verkar relativt hållbar i praktiken återstår att faktiskt finna de nämnda strukturerna, vilket är vad som beskrivs i följande avsnitt.

3.2 Rangbaserad chunkning

Att utgå från huvudsatsnivån, först med de primära begränsade leden, och därefter – delvis med hjälp av uteslutning – identifiera de obegränsade (rekursiva) primära leden i den 'luck-övning' som återstår, kräver inte desto mindre ett sätt att avgränsa de rekursiva leden i samma fält från varandra. Den metod som här redovisas kallas *rangbaserad chunkning* och har inledningsvis samma primära syften som chunkningsprocesserna i *Cass-Swe* (1998) och *Granska Text Analyzer, GTA* (Knutsson 2005). Men metoden har däremot inte som målsättning att parsa textmeningar till en frasstrukturell analys genomförd i någon av språkklasserna i Chomsky-hierarkin (jämför med *Cass-Swe* som utgår från reguljära uttryck, och *GTA* som utgår från något som liknar frasstrukturella omskrivningsregler). Rangbaserad chunkning har inte utvärderats enskilt, utan det resultat som denna metod ger ingår i den totala analysuppgiften.

Termen *chunk* introducerades av Abney (1991) och motiverades genom att dessa enheter skulle vara relaterade till ett huvudord och vara en användbar enhet från ett psykolingvistiskt och talspråksmässigt perspektiv (Megyesi 2002). En chunk innebär i litteraturen typiskt ett NP-huvudord med framförställda attribut, dvs. en nominalfras utan efterställda attribut. Eftersom prepositionen räknas som en prepositionsfras huvudord är en motsvarande PP-chunk istället en prepositionsfras fram till och med rektionens huvudordord om det är NP-format. En PP-chunk är därmed i praktiken detsamma som ett nominalt led (fast utan efterställda attribut för t.ex. NP) föregånget av en preposition.

Att, som här, nå de primära ledens *fulla* sträckning – inklusive efterställda attribut – är en annorlunda uppgift jämfört med chunkning, som alltså oftast inte inbegriper efterställda attribut. I fundamentposition är dock uppgiften enkel eftersom endast ett led i allmänhet ryms där. För att finna gränser mellan primära led i de andra fälten är uppgiften däremot ibland svårare än att identifiera NP/PP fram till dessas nominala huvudord. För att lyckas åstadkomma en analys i närheten av primär satslösning praktiseras den nedan nämnda chunkningsmetoden tillsammans matchning med hjälp av valensdata som är omarbetad från ordböcker, samt en stor uppsättning heuristiska regler för att även få med t.ex. efterställda attribut.

I den aktuella implementationen identifieras genom nedanstående segmenteringsmetod, *rangbaserad chunkning*, först, 1) en del NP-liknande typiskt nominala chunktyper vars speciella undertyp bestäms av huvudordet (rena NP, räkneordsfraser, adjektivfraser, participfraser), 2) *som*-fraser vilket är en benämning på ett segment av *som* (taggad som *konjunktion*) följt av NP-chunk (*som målvakt*) eller PP-chunkar (*som på medeltiden*), 3) adverbialchunkar (*lika bra*) och 4) PP-chunkar (*på medeltiden*). De tre senare har, om de är fristående led, grovt

sett adverbialfunktion. Till skillnad från flera *chunkare/shallow parsers* identifieras dock inte verbfraser eller liknande – chunkningen sker i de av de begränsade schemaleden avgränsade fälten i huvudsatsen.

Jämfört med en välformulerad uppställning av chunktyper hos Megyesis (2002) innebär segmenten att adverbfraser, adjektivfraser, numeriska uttryck, nominalfraser, prepositionsfraser och infinitivfraser har motsvarigheter hos chunktyperna i den aktuella metoden. Däremot saknas det som kallas verbkluster (’verbgrupp’, dvs. en sammanhängande sekvens av finita och icke-finita verb tillhörande samma verbfras t.ex. *skulle ha varit*). Verb utgör i den primära satslösningen här antingen, om de är primära, gränslinjer mellan de olika fälten vari segmenteringen sker, eller, när de licensierats, delar av attribut.

I SAG (1999) tydliggörs den svenska NP-strukturen med ett tiotal olika scheman. I nedanstående scheman (Tabell 20, Tabell 21 och Tabell 22) ska tolkningen vara att attributled är optionella.[41]

Definit attribut	Kvantitetsattribut	Adjektiv-attribut	Substantiv	Efterställda attribut
Dessa	*två*	*stora*	*böcker*	*om Hjo*
Mina	*många*	*andra*	*vänner*	*i Sala*

Tabell 20 Schema 1 i SAG (1999): ”Nominalfrasens vanliga ordföljd” (Band 3, s. 13)

Definit attribut	Deskriptivt attribut	Egennamn	Efterställda attribut
Den där	*lille*	*Albin*	*med sina många skrupler*

Tabell 21 Schema 3 i SAG: Del av ”Definit nominalfras med egennamn som huvudord” (Band 3, s. 36)

Emfatiskt attribut	Definit pronomen	Kvantitetsattribut	Adjektivattribut
Alla	*ni*	*båda*	*unga*

Tabell 22 Schema 5 i SAG: Del av ”Definit nominalfras med 1 och 2 ps-pron i plur” (Band 3, s. 38)

SAGs noggranna uppdelning i attributtyper är välanpassat för överskådligheten, men frågan är här om de olika attributtyperna har särskilt tydliga kopplingar till särskilda ordklasser. SAG kan sägas använda många scheman på grund av sina många olika syften och bl.a. görs en uppdelning mellan definita och indefinita

[41] De scheman som visas är bitar av de scheman som visas i SAG – där finns fler rader med exempel.

nominalfraser som inte verkar lika viktig i alla sammanhang, exempelvis inte för ren igenkänning.

Ex 32 NP → (Determinator) (Adv)* ({Adj, Räkneord, Particip})*
{Substantiv, Egennamn, Pron}

I generativ grammatikbeskrivning kan en beskrivning av NP utan efterställda attribut göras med mycket mindre specificitet men med bra uttryckskraft – en standardregel skulle kunna likna det reguljära uttrycket i Ex 32. Jämfört med de mycket större mekanismerna för att fånga nominalfraserna i de nämnda parsningsprogrammen ser sådana här beskrivningar enkla ut. Till det reguljära uttryckets fördel är att den använder ordklasser som brukar användas i ordklasstaggare. Nedan är några avgörande iakttagelser för den aktuella metoden.

- Det finns en typisk klass av *huvudord:* substantiv, pronomen, egennamn etc.
- Det finns en typisk klass av *bestämningsord* – determinatorer etc.
- Det finns en stor mängd *mellanord* – adjektiv, räkneord etc. – som kan fungera både som huvudord och bestämningsord.

Tendensen är att mellanorden (*mo*) inte avslutar en NP när ett ’huvudord’ (*ho*) följer, som i Ex 33.

Ex 33 *”glada/*mo *gäster/*ho*”*

Däremot är sekvensen *mellanord – bestämningsord* (*bo*) en indikation på NP-gräns, enligt Ex 34.

Ex 34 *”glada/*mo *ett/*bo*”*

Det går inte fullt ut att göra entydig lokalt grundad generell NP-chunking p.g.a. mellanorden, det blir i så fall en form av heuristik. Det nedanstående utgör ett bevis på att två identiska sekvenser av ord, *de två bilarna*, bör uppdelas olika beroende satskonstruktion, där den nedersta satsen kräver objekt. – Detta talar för att en metod som saknar valensinformation rentav *borde* ge mer än en analys. Hur många nominala led som ska/kan förväntas är viktig information i sådana här fall.

Ex 35	Idag	såldes	de/bo två/mo bilarna/ho . (En NP)
Ex 36	Idag	sålde	de/bo två/mo bilarna/ho . (NP 1) (NP 2)

Ex 35 och Ex 36 pekar på en typ av generell oavgörbarhet för segmentering utifrån ofullständig kontext som finns i svensk grammatik – det kan rimligen vara svårt även med andra parsningsmetoder.

Algoritm för den rangbaserade chunkningen

Till grund för den utvecklade algoritmen ligger ett försök till allmängiltig uppställning av i vilken ordning framförställda attribut förekommer. En första iakttagelse är det inre förhållandet mellan ordningen hos framförställda attribut och NP-huvudord, utgående från ordklasstaggningen i SUC 2.0.

Determinator	Possessiv	Grundtal Adverb	Particip Adjektiv Ordningstal	Substantiv Personligt pronomen Egennamn

Tabell 23 En första allmän beskrivning av ordningen hos ordkategorier i NP utan efterställda attribut ger följande uppställning av grupper utgående från de tidiga iakttagelserna, bl.a. SAG*s* scheman.

Utgående från Tabell 23 kan en beskrivning göras så att en NP-struktur, utan efterställt attribut, kan bildas genom att plocka ordklasser från vänster till höger på valfritt sätt från de olika grupperna. En beskrivning av en sådan NP-struktur är då en ordsekvens där ordklasserna följer denna regel, t.ex. possessiv – adjektiv – substantiv (*sin gamla bil*). En sekvens i en ordkedja som bryter mot ordningsuppställningen, t.ex. ordningstal – determinator – substantiv (*sju den bilen*) antas istället signalera att följdbrottet är en markör för ledskifte, här avslutas ett segment och en ny NP inleds med determinator.

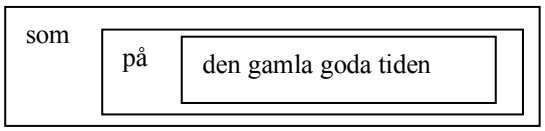
För de flesta ordgrupperna i Tabell 23 gäller att flera ord från en viss gruppering kan förekomma utan att frasen verkar avbrytas och ett nytt segment inleds, t.ex. determinator – adverb – adjektiv – adjektiv – substantiv (*Den mycket gamla röda bilen*). Däremot verkar sekvenser med ord från den högraste grupperingen ibland innebära en frasgräns (*bilen Johans, mig sten, blommor vi*) och ibland inte (*kilo mjöl, hertiginnan Beatrice, Karl Svensson*). Vad som saknas för en användbar hantering är för det första en klassificering av vilka substantivord som tar NP-attribut, och dessa kallas här *persontitlar* (*hertiginnan* i exemplet, kallas även *fast apposition, epitet*) och *mängdord* (*kilo* i exemplet).[42] Huvudregeln är här att en ordsekvens som innebär ett brott i den ’högergående’ kedjan *eller* två angränsade ord från den högraste gruppen markerar frasgräns. För att ändå inte markera persontitlar och mängdord följt av ord från den högraste gruppen som en ny segmentstart placeras de i en egen gruppering näst längst till höger. När

[42] Om tolkningen är sådan att ”måttsattributet” istället ses som huvudord kallas det efterföljande ordet istället *innehållsattribut*, se Ljung och Ohlander (1971).

det gäller att hantera *personnamn* rätt krävs listningar av för- och efternamn. När en sekvens verkar vara ett personnamn ska det inte innebära en frasgräns.

Utgående från denna beskrivning av NP utifrån dessa grupperingar kan nu undersökas hur förhållandet mellan dessa NP-liknande strukturer och två andra mycket viktiga strukturer ser ut. *Prepositionsfraser* har på flera håll enkelt beskrivits som en preposition följt av ett godtyckligt nominalt led. Det blir därför enkelt att lösa hanteringen av dessa i detta sammanhang genom att skapa en egen gruppering längst till vänster i Tabell 23. När ett segment inleds med en preposition betraktas ledet som en PP.[43]

Den sista kategorin i denna grupperingsbaserade metod är den syntaktiska struktur som inleds med *som* (eller möjligen *likt/såsom*) där ordet inte är relativbisatsinledare (*HP/HA* i SUC) utan konjunktion (*KN*). Detta segment har en funktionalitet och position som gör att den här ses som ett sorts adverbial och kallas *som-fras*.[44] Exempel är *som målvakt, som en helt ny komposition* eller *som i Schweiz*. Denna frastyp som även återfinns i Mamban kan i likhet med *som*-satser och PP fungera både som självständiga led (tolkas som adverbial) och som attribut. Som självständiga led betraktas de här som en adverbialtyp. Iakttagelsen här är att denna speciella ordgrupp tar NP – eller PP – som komplement och alltså bildar ytterligare en egen grupp längst till vänster i Tabell 23, och fungerar enligt Figur 8.



Figur 8 En insikt bakom den ranguppsättning som används är att PP-komplement (rektion) kan vara en NP och att den underordnade delen i vad som kallas en *som*-fras kan vara en PP (eller NP).

För praktiskt implementerat arbete har den likaledes schemabaserade beskrivningen utgående från Tabell 23 formulerats numeriskt, med hjälp av ranger. Efter en period av experimenterande har sammanställts en enkel uppsättning av rangtilldelning huvudsakligen beroende på ordklasstilldelningen i SUC, enligt Tabell 24. Det är alltså ett system som tilldelar frastypiska ordklasserna en konstant rang oavsett kontext.

[43] I detta sammanhang, den rangbaserade chunkningen, hanteras bara nominala led som är NP och närliggande frastyper och inte bisatser. Grundregeln att prepositionsinlett led är en PP gäller även där, men komplexare nominala led som t.ex. *att-satser* identifieras först senare i processen.

[44] I Wilhelmsson (2008) kallas denna strukturtyp *som-predikat*.

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

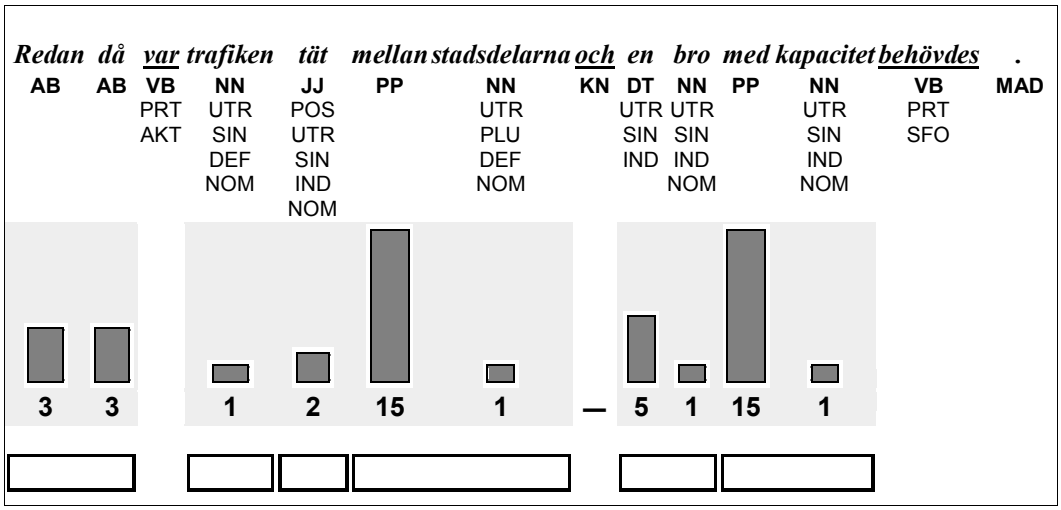
Tabell 24 Tabellen avgör många gränser mellan nominalfraser, prepositionsfraser och *som*-fras då en gräns känns igen i en genomlöpning som en övergång till en högre rang eller, i vissa fall, ett angränsande par av ord med samma rang. Ranger ges alltså även till adverb, konjunktioner etc., såvida dessa inte redan tillskrivits en roll som primära begränsade led, alltså som fristående adverbial eller primära konjunktioner.[45]

I uppställningen i Tabell 24 avgörs rangen således nästan enbart utgående från ord, ordklasstagg och listningar av måttsattribut, persontitlar, personnamn och mängdattribut. Detta innebär att många feltaggningar rörande morfosyntaktiska värden för genus, bestämdhet, numerus, diates m.m. ignoreras. På detta sätt liknar därmed detta den analys som görs i inledande sammanfogning av frassegment före kongruensundersökning i flera program för grammatikkontroll. Dvs. det liknar en *vid* eller *tillåtande* grammatik. Med relativt få undantag har denna metodik visat sig fungera väl för det aktuella syftet. Tabell 25 och Figur 9 visar hur rangerna identifierar konstituenter.

[45] I Wilhelmsson (2008) har *Persontitel* med rangen 1,5 fallit bort från uppställningen. Hoppet i rangvärde från 5 till 14 beror på att värdena är satta som en kortleksanalogi. 1-5 svarar mot de centrala ordklasserna i NP.

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		*någ-ra*		*bra*		*koppar*	*kaffe*	PP
				Sitt		*röda*		*hus*	NP
								Per Karlson	NP
					Ganska bra				Adverb
			De			*första*			NP

Tabell 25 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.



Figur 9 En illustration av ranger som stapellängd visar hur en högre stapel än den till vänster i ordsekvensen indikerar ny chunkstart (hb06a-011). De begränsade primära leden är understrukna.

Rangbaserad chunkning har som främsta uppgift att finna gränser mellan rekursiva *('unbounded', obegränsade*) led i svenska meningar i området som kommer efter primärt finit verb (dvs. mitt- och efterfält eller 'efterdel'). Den algoritm som använder detta rangsystem kan beskrivas som nedan.

1. Områdena där de aktuella strukturerna ska identifieras behandlas först så att varje löpord med ordklass får en rang enligt ovan, som i Figur 10.

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 10 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 siffra) eller lika rang som det föregående för att tolkas som del av samma segment (chunk).

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

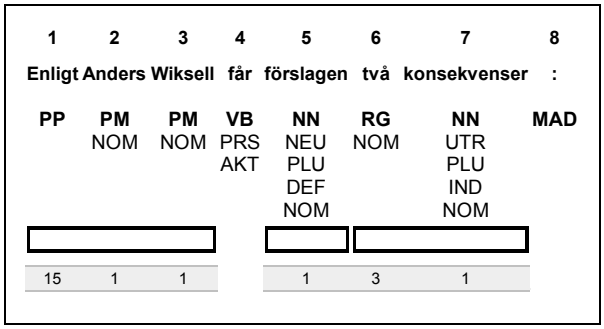
   – Detsamma gäller när två ord av rang 1 finns bredvid varandra – men där görs en speciell undersökning så att fulla personnamn *Bea (1) Karlsson (1)* räknas som samma struktur om båda är taggade som egennamn (*PM* i SUC) och det första ordet finns i en listning av förnamn och/eller det andra ordet finns 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 11 Efter rangtilldelningen skapas inledande chunksegment.

I Figur 11 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.[46]

- 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.

Som framgår av tabellen med ranger kräver metoden, utöver ren tagginformation, listningar av ord med speciell funktion. Den speciella rangen *1,5* indikerar att motsvarande ordklass för denna ordgrupp normalt är 1[47] men att dessa grupperingar av undantag alltså kan ta komplement. (Det är dock inte alla ord med rang 1 som fungerar som komplement, t.ex. inte pronomen.) Om detta komplement inleds med högre rang än 1, t.ex. *ett/3 kilo/1,5 mogna/2 äpplen/1* utgör räckan av ranger dock en bruten kedja, dvs. analysen ser *ett kilo* och *mogna äpplen* ut som två skilda segment. Detta är faktiskt meningen – sammanfogningen till en chunk sker i ett senare skede där det klargjorts hur många nominala led (dvs. subjekt plus objekt/predikativ) som eftersöks, eftersom det *kan* vara frågan om två skilda chunkar. Tolkningen bör idealiskt ta verbvalens i beaktande. Ex-

[46] 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.

[47] Dessa ord är nästan uteslutande substantiv, med ett fåtal undantag som *Ångström (*personnamn*)*. Ord som markeras med rangen 1,5 kan inte ta vilken typ av komplement som helst: efterföljande pronomen är t.ex. ett undantag.

empel på detta är *Idag mötte tennisspelaren/1,5 hårtslående/2 Karlsson/1* där det monotransitiva *mötte* kräver objekt, och chunkarna inte bör sammanfogas av denna anledning. Detta kan jämföras med fallet *Lendl mötte tennisspelaren/1,5 hårtslående/2 Karlsson/1*, där bara ett nominalt led förväntas efter finitet. Metoden är gjord för att kunna skjuta upp somliga beslut till ett senare läge där funktionell etikettering görs.

Ex 37 *abbé, ackompanjatör, actionstjärnan, adelsmannen, adjunkt, adjunkten, adress, advokat, advokaten, affärsmagasinet, affärsmannen, affärstidningen…*

I skrivande stund innehåller implementationen drygt tusen persontitlar. Ex 37 visar de första. I likhet med andra listningar som för anföringsverb och mängdord skulle komplett listning vara omöjligt. Dessutom har utifrån dessa byggts en suffixgrundad funktion för att låta icke införda ord, med t.ex. suffixet *-inna* fungera på samma vis (dvs. få rangen 1,5). Denna grupp innehåller nu inte enbart rena persontitlar utan en del ytterligare ord som har samma funktion, t.ex. förstadelar i tvådelade platsnamn (*St., Los, Rue*).

Ex 38 *aln, ampere, andel, andelen, antal, antalet, appendix, arealen, armada, armé, arsenal, art., artikel, ask, avd, avdelning, avdelningen, avsnitt…*

I skrivande stund innehåller implementationen drygt 500 mängdord (*measure nouns*) varav några visas ovan. I somliga fall som *10-tal, 20-tal, 30-tal* etc. speglar samlingen snarast den mängd i ordgruppen som råkat påträffas i träningsmängden av SUC 2.0. Mot denna metodik som särskiljer ordgrupper i detta skede av parsningen, kan ställas alternativet att göra en modifiering av SUCs taggset från början i ordklasstaggningen, med tillägg av specialtaggar för mängdord, vissa hjälpverb tidsuttryck etc., vilket då i vissa aspekter också kan förbättra korrektheten för ordklasstaggningen med en sådan tagguppsättning (Forsbom 2008).

I fråga om *personnamn* används listor på dryga 12 000 och 9 000 för- respektive efternamn.[48] Personnamnen är företrädesvis svenska sådana men har utökats, vilket kan göras direkt från programmets gränssnitt. Tyvärr insamlades även personnamn och etiketterades som för- eller efternamn från hela SUC, dvs. även från den i övrigt orörda testmängden. Det betyder att det finns en liten risk att denna insamling otillbörligt har gynnat utvärderingarna med testmängden.

Ett sent tillägg till algoritmen har varit hanteringen av utländska ord (*UO*, i SUCs taggningskoder). Dessa får i programmet en heuristisk hantering som nominala huvudord (de får rang 1 i rangchunkningen) vilket oftast leder analysen rätt, eftersom det ofta är fråga om en metaspråklig hantering oberoende av

[48] Dimitrios Kokkinakis tackas för dessa.

grundbetydelsen (ordklass på originalspråk) som i Ex 39 där dessa ord är understrukna.[49]

Ex 39	a) *Epinikion*, bokstavligen ”till segern”, är en lovsång […]	(ja21-057)
	b) […] av rapporter , som ger ”*a true and fair view*”.	(jc07-014)
	c) ”Elativen *pojasta* i satstyp (4b) har av vissa grammatiker […]	(ja12-072)

3.3 Rangbaserad chunkning i jämförelse med några andra typer av chunkning

Det följande är en kort jämförande beskrivning av några andra chunkningssystem för svenska som också ibland kallats *shallow parsers*. Systemen, som representerar olika ansatser för den relativt likartade uppgiften, är *MorP, GTA* (*Granska Text Analyzer*) och *Cass-Swe*.[50] Syftet är att visa vilka kvaliteter fungerande system har.

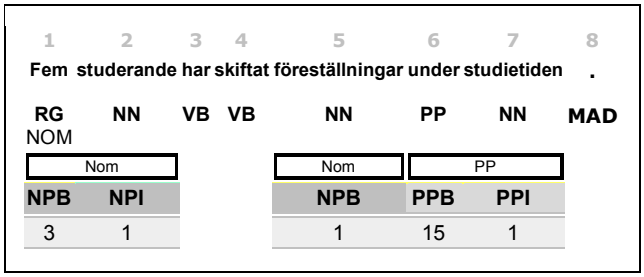
3.3.1 NP-identifikation i system med ytstrukturanalys

Systemet *MorP* (Källgren 1992) använder ett mycket litet lexikon och speciella suffixregler för ordklasstaggning. Regelsystemet har inte undersökts, i rapporten finns bara vissa delar. (*MorP* finns möjligen tillgängligt på Stockholms universitet men har inte testats.) En skillnad mellan rangbaserad chunkning och *MorPs* chunkning är den Constraint Grammar-liknande uppsättningen av flera möjliga ordklasstaggar per ord i *MorP*, vilket verkar leda till en komplicerad tolkningsprocess ibland. Exempelvis betyder den ’icke disambiguerade ordklassen’ *W:* ’substantiv *eller* verb’, och NP-identifikation är till skillnad från här första steget

[49] De senare sammanfogningsreglerna innehåller en regel som sammanfogar angränsande UO-taggade ord. Det innebär att ett segment av utländska ord inte kan innehålla satsledgräns.

[50] Dimitrios Kokkinakis (*Cass-Swe*), Ola Knutsson och Jonas Sjöbergh (*GTA*) tackas. De har för nedanstående undersökningar varit mycket tillmötesgående i korrespondens och konversation. Hela NP-regelsamlingen i *Cass-Swe* och hela *GTA*-grammatiken och ett speciellt Python-program för att använda *GTA* över Internet erhölls därför. En detaljerad undersökning angående vilka former av NP som *GTA* och *Cass-Swe* identifierar precist har gjorts.

efter denna taggning.[51] Att primära finita och infinita verb identifieras i steget före NP-identifikationen i föreliggande program leder till att NP-identifikationen från början sker på områden avgränsade av primära begränsade led. Annars är likheten att båda metoder känner igen NP (i första steget utan efterställda attribut) genom att *gränser* och *icke-ingående ordklasser* känns igen. De båda metoderna är därför eventuellt exempel på *distuent-grammatiker* för svenska.[52] De är dock inte statistiskt grundade, vilket är fallet i Magerman och Marcus (1990) för engelska. I implementationen i detta arbete är den s.k. IOB-notationen en valbar information som kan visas som i Figur 12.



Figur 12 Resultatet från rangbaserad chunkning kan tydligt beskrivas med s.k. *IOB-markering* från Ramshaw och Marcus (1995), dvs. markering om huruvida varje ord är en segmentstart (*B, Beginning*), inom segment (*I, Inner*) eller utom ett segment (*O, Outside*). I denna implementation kan sådan markering optionellt visas i analysen, som här ovanför rangraden, även om den inte tillför någon information. (jb05-063)[53]

[51] Tabellen nedan visar olika ordklassers roller i nominalfraser i *MorP* (Källgren 1992).

Ordklasser som både kan inleda och finnas i NP	*NP-avslutande ordklasser ("stoppare")*
J Adjektiv	Alla verbslag
O Possessivt pronomen	Konjunktioner
Q Kvantifierare	Skiljetecken
X Adjektiv / Verb	
Y Adjektiv / Substantiv	

Ordklass som ibland avslutar men inte inleder NP	*Huvudord – förutsatt att det kongruerar med föregående ord, annars: föregående NP avslutas och detta är egen NP*
W Substantiv / verb	N Substantiv

Tabell: Ordklasstaggningen i *MorP* ger en speciell uppsättning ordklasser. "Stopparna" är sådana ordgrupper som betyder att tidigare pågående NP därmed har avslutats.

[52] Denna ovanliga term verkar användas för grammatiker där avgränsning har betydelse i chunkning till skillnad från enbart matchning, vilket i så fall är en passande beskrivning av den aktuella metoden.

[53] De figurer visande strukturanalys som finns i detta kapitel kan skapas i implementationen som ett alternativ till HTML med färgkodsmärkning. Etiketterna på chunkarna under ordklasstaggningen här är ibland förändrade (förtydligade) för hand.

I arbetet med *Cass-Swe*, som är en svensk version av Abneys arbete (1997), har tidigare försökts att sammanfoga efterställda attribut till nominalfraser (uppgiften *PP-attachment*). Detta gäller dock inte för systemet som används idag. Dessutom är reglerna byggda med hänsyn till feltaggning inom systemet, vilket bidrar till det stora antalet regler.

GTA täcker nu (eventuellt redan vid tidpunkten för utvärderingen, se Tabell 26) *vissa* efterställda attribut, dock ej PP. Detta designval delar *GTA* med *Cass-Swe*. *CassSwe* använder numera *TnT* (Brants 2000) för taggningen medan Granska har använt en egen taggare som har snarlik korrekthet (*TnT* har emellertid också använts bl.a. i utvärderingen som rapporteras i Tabell 26).

	MorP	GTA (Granska)	CassSwe
Antal regler	Oklart. Några centrala behandlas i rapporten. Reguljära uttryck verkar användas mycket mer ekonomiskt än i *CassSwe*.	260 i hela grammatiken (inkl satsgränsregler)	> 700 enbart för NP utan efterställt attribut
Rapporterad korrekthet av upphovsmännen	Ej utvärderat.[54]	F-score: 91,4 % (Bigert, Knutsson och Sjöbergh 2003). Ca 15 000 ord från SUC (sex slumpvis valda texter à 2000 ord)	*CassSwe* har utvärderats för NP-identifikation mot medicinsk text med F-score: ca 96 %. Då räknas ej halvträffar' som rätt.
Formalism/relation till rangsystem	MorP avgränsar NP bl.a. genom att känna igen typiska NP-starter och ord som ej kan ingå (verb etc.). Detta är ganska likt ett rangsystem. Reguljära uttryck.	Egen sofistikerad formalism med omskrivningsregler. Ett system av matchningsregler och hjälpregler.	Huvudsakligen reguljära uttryck (ev. med inslag av starkare formalism). 'Longest match' (ett tjugotal samlingar för NP av sjunkande komplexitet). De mest komplicerade provas först etc.

Tabell 26 Chunkningsmetoder för svenska i tre olika tidigare ansatser.

[54] Nomenigenkänning hade precisonen 95,1 % och recall-värdet 90,5 %. Denna lägre korrekthet hos taggningen borde ha givit sämre NP-korrekthet. Detta säger inget om systemets korrekthet med korrekt eller likvärdig taggning.

Resultaten i Tabell 26 verkar inte fullt kunna jämföras p.g.a. olika definitioner och texter. Efterställda attribut ingår ej i uppgiften (möjligen delvis i GTA). En undersökning om vilka faktiska NP-konstruktioner som systemen täcker har genomförts (opublicerad) i samband med detta. Generellt kan kanske sägas att chunkningen har en mer uppbärande roll för de andra systemen, där den utgör första steg, och där satsnivåer ej klargörs som görs här med identifikation av de begränsade leden på huvudsatsnivå.

3.3.2 Svagheter och svårigheter med rangbaserad chunkning

Rangbaserad chunkning har följande positiva och negativa egenskaper som delvis skiljer ut den jämfört med andra metoder för denna deluppgift i parsningen. De två sista punkterna innebär krav som metoden ställer.

- Metoden kan som här göras ’transparent’, med uppvisande av rangerna, under körning (därmed kan fel beroende på taggning spåras).
- Den identifierar minimala NP (med olika undertyper), PP, AP, AdvP samt *som*-fras. Chunkens högsta rang (lägsta siffra) markerar generellt dess huvudord och typ, med undantag för PP och *som*-fras.
- Den kräver inte användning av en ”generativ grammatikformalism” som omskrivningsregler eller reguljära uttryck för att programmeras.
- Som alla svenska system riktar den inte in sig på täckning av efterställda attribut direkt, dvs. egentligen inte på full frasidentifikation. Tillsammans med de efterföljande segmenteringsmetoderna har metoden dock detta syfte.
- Syftet är att fungera tillsammans med efterföljande regler där en del sammanfogningar är tänkta att ske för att nå det slutliga syftet att skapa NP och med ’maximal projektion’, dvs. de nominala och adverbiella sjok som utgör kandidater till obegränsade primära funktionella led (huvudsatsernas subjekt, objekt/predikativ och adverbiella led med deras fulla sträckning).
- Den är i nuvarande version beroende av korrekt ordklasstaggning men den är i många fall okänslig mot fel i ingående särdragsvärden utöver ordklasser och undersöker inte kongruens.
- Den kräver utarbetade listor av ordmängder (egennamn etc.) och, för sammanhanget här, en föregående licensieringsprocedur. För att avgränsa de fält där chunkningen sker krävs också listning av vissa ordgrupper (fristående adverbial) som licensieras eller är primära och som inte enbart kan kännas igen genom taggningen i SUC 2.0.

Det finns några inneboende svagheter med att använda detta rangsystem för chunkning, det handlar oftast om att den ordklassindelning som finns i SUC (som alltså används) faktiskt inte är tillräckligt specifik. Ett typiskt exempel är

adverb (rang 3) som enligt målsättningen bör behandlas på minst tre olika sätt beroende på olika syntaktiska egenskaper.

- En klass mittfältsadverb, t.ex. *inte,* ska i allmänhet bli ett eget led (dvs. inte ingå i annan struktur) och de adverb som potentiellt fungerar så måste listas specifikt.
- En annan klass adverb är potentiellt pre-modifierande och kan i vissa fall ses modifiera verb, dvs. *fokuserare*, t.ex. *nästan och kanske*. *Han nästan sprang.*
- Ytterligare en annan klass adverb fungerar postmodifierande – *adverbattribut*. *Vägen hem, resan dit* etc.

Denna uppdelning av adverben finns inte i ordklass- och särdragsinformationen som ingår i SUC 2.0 (eller i ordklasstaggare som bygger på SUC). Istället måste listor i vissa fall sammanställas för att chunkningen enligt den beskrivna metoden ska blir rätt.

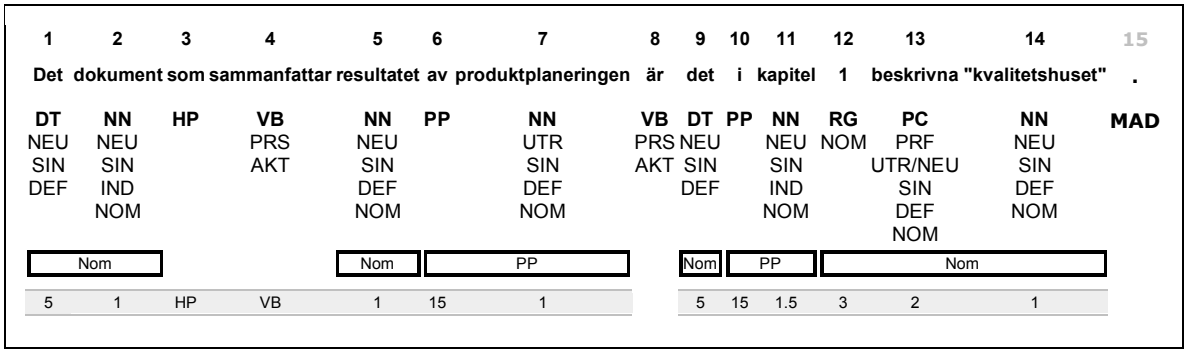
En annan aspekt är att en prepositionsfras som följer på en annan fras generellt inte kan bestämmas (förutom i fundamentposition) som attribut eller fristående utan vidare undersökning. Denna fråga lämnas istället till en senare procedur i analysen, där länkningar från egna tillägg och två valenslexikon används.

Syftet här har varit att hålla metodiken så enkel som möjligt och att inte modifiera den ovannämnda samlingen av krav hos indata. Den innebär dock en metod som från början inte är perfekt – en del undantag finns. Dessa undantag demonstreras nedan.

Metoden fungerar i praktiken med bra resultat när det gäller att identifiera fraser fram till huvudordet (i PP/*som*-fras fram till rektionens huvudord). De *felaktiga identifikationer* som görs kan alltså uppdelas i 1) *ej helt täckta egentliga chunkar* och 2) *markering av längre segment än de egentliga strukturer som borde identifieras* (felaktig sammanfogning). Medan ej helt identifierade chunkar kan sammanfogas till de rätta strukturerna i senare sammanfogningssteg, på samma sätt som t.ex. efterställda attribut, så är felaktiga sammanfogningar av ord i rangchunkningen vanskligare, men dessbättre mycket mer sällsynt.

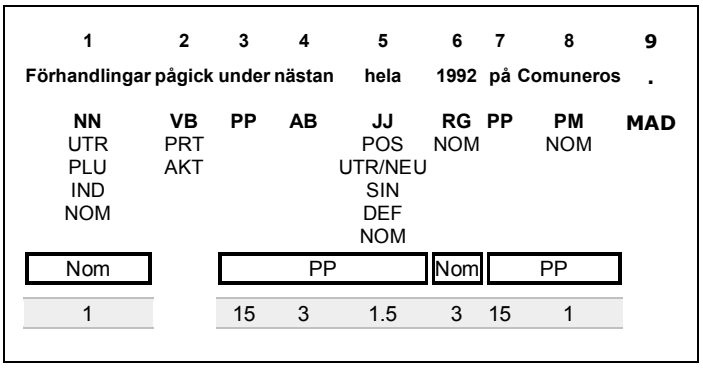
Ej helt täckta segment

Det vanligaste slaget av ej helt riktigt täckta chunkar är inskott av prepositionsfraser bland räckan av framförställda attribut i nominal- och prepositionsfraser som i Figur 13.



Figur 13 Inskott av prepositionsfras i räckan av framförställda attribut som här *det i kapitel 1 beskrivna "kvalitetshuset"* är ett någorlunda vanligt undantag som innebär att en egentlig fras missvisande delas upp med hjälp av rangchunkningen (je01-026). Det egentliga ledet *resultatet av produktplaneringen* ses däremot inte som felaktigt uppdelat: Den rangbaserade algoritmen identifierar ju inte efterställda attribut utan lämnar dessa sammanfogningar till senare steg. (I detta fall sammanfogas detta segment genom en i programmet beskriven ordlänkning mellan *resultatet* och *av*, men dessutom p.g.a. att hela fundamentledet ändå sammanfogas).

Att egentliga segment inte alltid täcks som i Figur 13 och i Figur 14 är som nämnts meningen, och det finns i programmet många regler och en funktion i gränssnittet för att skapa sammanfogningsregler på grundval av ord och ordklasser i angränsande segment.



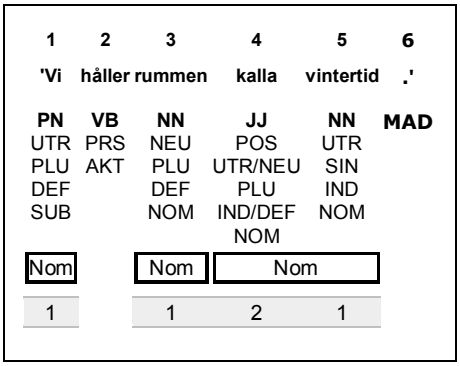
Figur 14 Årtalet 1992 har en funktion som ett NP-huvudord men får enligt rangtilldelningen värdet 3. Detta är en typisk konsekvens av rangerna som dock kan rättas till i följande segmenteringssteg. Sammanlänkningsregler kan skapas direkt i gränssnittet. (ec15d-006)

Felaktigt sammanfogade segment

Det är enkelt att komma på exempel på två angränsande segment som metoden som den utformats felaktigt sammanfogar. I praktiken är alltså dessa fall sällsynta (som i det redan visade i Ex 40) och oftast möjliga att undvika.

Ex 40 Att smyga sig in på en kronhjort är något av det svåraste jägarlivet har att bjuda på. (ea25-125)
15 5 2 1

Ex 40 exemplifierar ett genuint svåravgränsat PP/NP-par som inte heller kan skiljas isär med hjälp av särdragsinformation, vilket används flitigt i andra system. *Av det svåraste jägarlivet* är i sig en giltig PP-struktur med kongruerande ord. Att det är frågan om två segment (och en *som*-strykning) är sannolikt inte denna metod ensam om att missa. Det kan verka som ett riskabelt projekt att inte arbeta med särdragsvärden och kongruenskontroll. Det ger emellertid förvånansvärt sällan upphov till fel. Figur 15 visar ett fall av felsammanslagning där *kalla* av misstag sammanfogas med NP (tidsadverbial). *Kalla vintertid* är i sig en giltig NP.



Figur 15 Felaktig sammanfogning sker som här förhållandevis sällan p.g.a. rangerna. (ja12-024)

3.4 Stegvis sammanfogning av chunksegment till större enheter

Efter den rangbaserade chunkningen sker en stegvis sammanfogning av dessa segment till allt större enheter med syftet att möta ett antal primära nominala led som är kompatibelt med den aktuella sats- eller verbfrasstrukturens behov, utgående bl.a. från dess valens. I dessa steg sammanfogas också primära adverbiella led av ett antal som är okänt och teoretiskt sett obegränsat. Typerna av segment som skapas är också fler än de som är resultatet från den rangbaserade chunkningen, eftersom den stegvisa sammanfogningen skapar komplexa typer som bisatsled och infinitivfraser. Dessutom har detta steg syftet att fästa efterställda attribut till föregående segment. Målet är att de segmenttyper som i Telemans uppställning kan ha nominal roll (enligt Tabell 17) och dessutom adverbiella segment ska identifieras, för att utgöra kandidater i den funktionella rolltilldelningen, enligt Figur 7. I många fall är redan resultatet från den rangbaserade chunkningen tillräckligt och inga fler sammanfogningar sker. Men komplexare enheter bestående av flera potentiellt obegränsade strukturer som infinitivfraser

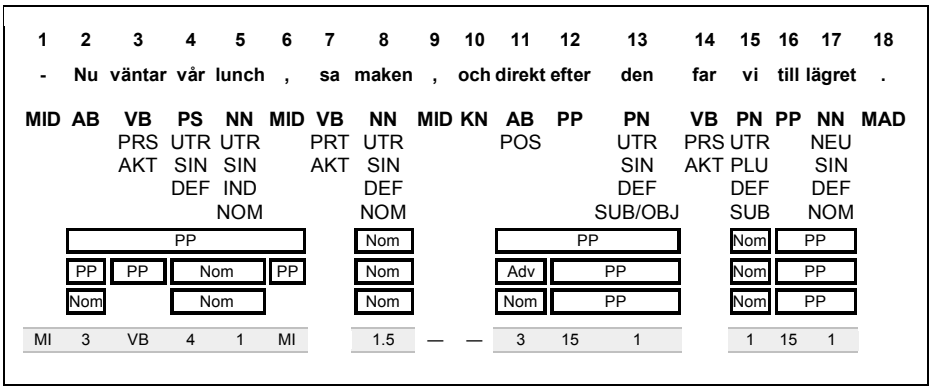
och bisatser skapas t.ex. genom dessa sammanfogningsprocedurer. Viss annan information om möjlig adverbialfunktion hos NP, troligt huvudord etc. läggs också till.

Den stegvisa utökningen av chunksegmenten till större enheter är en del av denna metod, som består främst av sammanlänkning av två eller flera chunkar på grundval av två angränsande ordtyper och/eller ordklassinformation. Detta är också ett steg där sådana möjliga sammanlänkningar ibland *inte* bör ske p.g.a. att satsen kräver att ett av leden ska inleda en ny chunk. Det är alltså ett av de lägen där möjlig strukturell ambiguitet som felkälla tydligt uppenbaras och i princip kan hindras.

3.4.1 Framförställda attributslag

Sammanlänkning av ytterligare framförställda attribut sker genom olika sammanlänkningsregler. De följande är typiska exempel.

- Sammanfogning av framförställda adverbattribut till prepositionsfraser krävs eftersom adverb har rangen 3 och prepositioner 15, vilket visas i Figur 16.



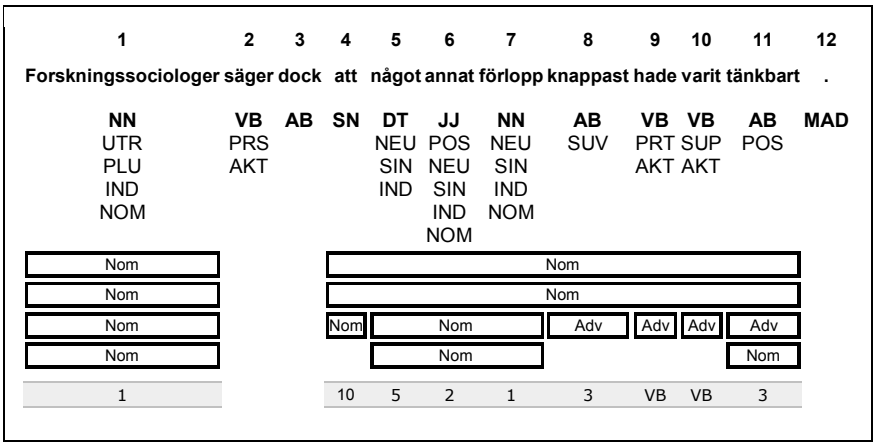
Figur 16 Sammanfogning av framförställda attribut sker här exempelvis med en regel *'direkt + preposition' (direkt efter den)* i det tredje steget (lagret) av sammanfogningar. Ovanstående exempel betraktas dessutom som felaktig analys, anföringen fortsätter efter *"sa maken"*. *Sa* borde tolkas som det enda primära finitet. (kn05-074)

- Vissa kvantifikatorer och mängdord har enligt rangmetoden inte länkats samman med följande delar. En anledning till detta är, som redan nämnts, att skjuta upp somliga beslut om hopslagning till senare i analysen då antalet tillgängliga led kan jämföras för matchning med verbställighet och möjliga komplementslag. Speciella kvantifikatorer som *halva* (adjektiv) kan föregå t.ex. determinator, *halva/2 den/5 stora vägen*.

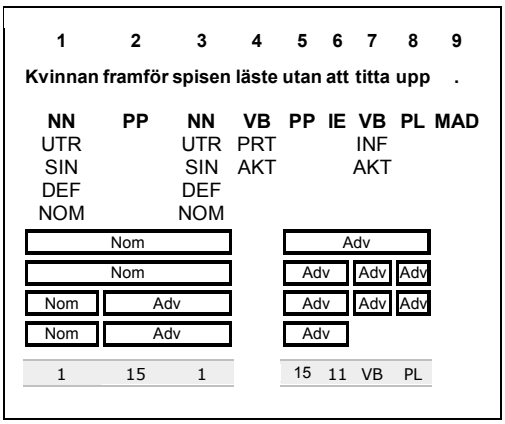
3.4.2 Efterställda attributslag

- Sammanfogning av efterställda PP-attribut på grundval av valens hos substantiv/adjektiv/particip i NP/PP utgående från valensdata från NEO*, Lexin* eller egna tillagda länkningar: *ackvisition – för, härskare – över, protest – mot*.

- Sammanfogning av efterställda relativbisatser till föregående struktur. En kvantitativt försvarbar (men långt ifrån undantagslös) tumregel är att påhängda relativbisatser och andra bisatser *inte efterföljs av andra primära obegränsade led i samma satsfält*. Detta exemplifieras i Figur 17. En liknande strategi för infinitivfraser visas i Figur 18.



Figur 17 Sammanlänkning av hela *att*-satsen sker till nästa begränsade primära led (fc01-090). De olika etiketterna *Nom* och *Adv* är defaultvärden som har mindre betydelse, speciellt för de lägre segmenteringsnivåerna. Nästa avsnitt redogör för hur de tre skikten av segment, under texten och ovanför det nedersta (rangchunkningsresultatet), skapas.



Figur 18 Satsförkortningstypen infinitivfras sammanfogas på samma sätt som i Figur 17. En annan regel innebär att en preposition (*utan*) tar infinitivkomplement. (kl16-154)

3.4.3 Tre lager av sammanfogningsregler

För att skapa denna funktionalitet som slår samman närliggande chunkar har ett kontinuerligt testande pågått med träningsmängden för att finna så många exempel som möjligt, varifrån reglerna, som är uttryckta så generellt som möjligt för att täcka ett visst fall, hämtats.[55] Sammanfogningen sker som syns i exemplen organiserat stegvis i tre skikt. Denna uppdelning möjliggör att finna felaktigheter och knyta dem till en specifik sammanfogningsprocess. De tre stegen har bara detta syfte och det har inte någon speciell betydelse vad som sker i vilket steg.[56] Framförallt ska det påpekas att sammanfogningen av segment inte motsvarar en konstituentstruktur eller liknande.

Strukturnivå 1

På strukturnivå 1 skapas utöver segmenten som kommer från den rangbaserade chunkningen egna chunkar för begränsade led, licensierade begränsade led, bisatsinledare, interpunktionstecken och interjektioner. I princip ska alla grafiska ord utom de identifierade primära begränsade leden, inklusive förfält, ingå i någon chunk under segmenteringen.

Strukturnivå 2

Sammanfogning av chunkar till större chunkar i strukturnivå 1 på grundval av:

– Ingående bisatser (inklusive relativbisatser) som har bisatsinledare eller saknar bisatsinledare (p.g.a. *som*-strykning) enligt föregående avsnitt
– Ingående infinitivfraser
– Ingående valensmatchning för substantiv och adjektiv/particip med PP

[55] Rent implementationstekniskt sparas s-enheter som exempel tillsammans med kommentar och flera exempel kan därmed jämföras för att skapa så exakta matchningsregler som möjligt senare (se Kapitel 4). Funktionalitet för att spara exempelenheter har allmänt använts mycket i implementationen, bl.a. för att spara enheter där feltaggning i SUC misstänks.

[56] En möjlig konsekvens av vilken ordning som stegen sker i är dock att om stora sjok sammanfogas tidigt i processen så kommer mindre parvisa sammanfogningar p.g.a. attributvalens och liknande att behöva ske mer sällan, eftersom de nedan ingår i sådana större strukturer.

Bisatser och infinitivfraser

Bisatser, inklusive relativbisatser, och infinitivfraser har naturligt en underordnad roll i huvudsatsanalysen eftersom de ingår i eller utgör egna primära satsled, men sällan behöver skärskådas i fråga om beståndsdelar för huvudsatsanalysens skull. I fall av utbrutna satsled, satsflätor etc., är det egentligen nödvändigt med denna analys av underordnade konstituenter för kunna etikettera led rätt. Denna aspekt har visat sig möjlig att bortse från i huvudsatsanalysen om målsättningen är att nå hög procent korrekta analyser, varifrån undantag senare kan särbehandlas, snarare än att den utgör en perfekt regelsamling/grammatikbeskrivning från början. Bisatser, inklusive relativsatser, och infinitivfraser har visat sig ofta underkasta sig följande enkla, men långt ifrån undantagslösa, heuristiska regel i parsningen.

- En inledd bisats eller infinitivfras tenderar att låta det primära led det utgör eller ingår i att fortsätta fram till nästa primära *begränsade* led (verb, partikel, reflexiva pronomen, korta adverbial) eller slutet på det primära fältet/den aktuella huvudsatsen. D.v.s: när ett obegränsat led (nominalt eller adverbiellt) innehåller (ev. relativ) eller är bisats eller infinitivfras, förväntas detta primära led att inte efterföljas direkt av ett annat primärt obegränsat satsled. Ex 41 visar några faktiska undantag från denna regel.

  Ex 41 a) Under den pågående vattenfestivalen är, som vi tidigare berättat, också kooperationen engagerad. *(som/HA-adverbial följt av subjekt etc.)* (he07c-002)
  b) Ingen kan veta vad jag drömmer, även om vi med EKG kan sluta oss till att jag drömmer. *(vad/HP-sats följd av subjunktionsinledd bisats)* (gb10-041)

Att kunna undvika att för långt segment ses som del av inledd bisats är ibland en svårare uppgift. I Ex 41 finns ledtrådar i a) genom att om tolkningen är att *som*-satsen sträcker sig ända till slutet så saknas subjekt och objekt. I b) krävs inga fler led om *vad*-satsen skulle fortsätta till slutet. En tendens är dock att subjunktionsinledd bisats (*även om*) är 'starkare' än relativsatser och därför blir ett nytt primärt led trots inledd underordnad sats. Dessa undantagsfall täcks inte riktigt bra ännu.

Strukturnivå 3

Sammanfogning av chunkar från strukturnivå 2 till större chunkar på grundval av:

– En stor samling av heuristiska regler som för samman chunkar p.g.a. deras innehåll, starter, avslut och ingående ord och ordklassinformation. Skapandet av olika länkningsregler kan ofta ske direkt i gränssnittet. Möjligheten att införa ord

i speciella grupper direkt vid en påträffad förekomst i speciell textmening innebär att samma s-enhet direkt kan testas efteråt. Ofta är en sådan ordkategorisering, t.ex. tillägg av mängdord, och inte nya allmänna syntaktiska regler, vad som krävs för korrekt syntaxanalys.

Dessutom sker sammanfogning på grundval av:

– Identifikation av parentesuttryck
– Prepositioner med speciella rektioner (bisatser och infinitivfraser)
– Konjunktioner i slutet av segment

3.5 Identifikation av primära subjekt, objekt/predikat och adverbial

När de obegränsade segmenten sammanfogats genom chunkningen och den följande, ovan beskrivna, segmenteringen har gjorts representeras den primära nivån på en linjär form exempelvis som ’nom finit nom refl adv infinit adv’. Uppgiften att tilldela funktionella syntaktiska roller består av en sorts kandidatinsamling och för de nominala leden, som är begränsade till antalet per sats, en räkning av dessa. Adverbialled, vilka alltså kan kännas igen direkt i de flesta fall som PP och AdvP, förekommer dock i helt okänt antal per sats.

3.5.1 Identifikation av primära subjekt och av primära objekt/predikativ

Etiketteringen av subjekt i denna ansats och implementation har tydlig uppdelning i olika konstruktionsfall, eftersom förutsättningarna för subjektsidentifikation skiftar kraftigt med olika satstyper och huvudverb. Identifikationen av objekt/predikativ är sammanflätad med subjektsidentifikationen: allmänt gäller att de nominala led som återstår efter subjektsidentifikation blir objekt. I likhet med identifikationen av begränsade led inleder ett steg där kandidater insamlas från de olika fälten. Subjekt brukar enligt de flesta uppskattningar i litteraturen vara fundamentled i minst 60 % av en texts huvudsatser men det finns en stor variation.

För avgörandet av vilket led som är subjekt och vilket som är objekt/predikativ kan beaktas tre aspekter i denna ansats, här utan inbördes ordning: 1) den konstruktionsmässiga/positionsmässiga ställning som leden befinner sig i, 2) de interna strukturmässiga/ordklassmässiga skillnaderna mellan leden och slutligen 3) de betydelsemässiga aspekterna hos leden, som från datorperspektivet måste fås genom listningar så att t.ex. de huvudord som har en animat referens (som inte framgår av själva den morfo-syntaktiska ordklasstaggningen) kan skiljas från andra.

De flesta faktiska förekomster i svensk text kan lösas korrekt genom beaktande av aspekterna 1) och 2). Dessa aspekter har också huvudfokus hos Sköld (1966). Hans uppställning över ledtrådar för subjektsidentifikation kan sammanfattas som i Tabell 27 (här omgrupperad).

Ledtråd	Exempel	Typ
Kasusmärkning hos pronomen klargör dess roll som subjekt eller objekt	*Den tar jag (SUBJ).* *Dig (OBJ) ser alla.*	Struktur/ordklass
Possessivt reflexivpronomen inleder objektled	*Sin flicka (OBJ) älskade sjömannen*	Struktur/ordklass
Det nominala led som ej är i fundamentet kan tolkas rätt genom positionsförhållande till förekommande nexusadverbial, verbpartikel eller hjälpverbskonstruktion	*Pekka fäller trädet (SUBJ) inte.* *Planet (SUBJ) sköt ned hindret.* *Pekka (SUBJ) har fällt trädet*	Konstruktion/position (med andra ord: satsschemats regler för leden placering i modern danska/svenska)
Subjekt–verbkongruens i numerus (ej längre användbar i svenska)	*På ifrågavarande ö Scandia bo många folk, men blott sju av dem (OBJ) nämner Ptolemaios.*	Struktur/ordklass
Prosodisk information (ej användbar i skrift)	*Kalle (OBJ?) träffade Lisa*	Talspråklig

Tabell 27 De tre första typerna av ledtrådar i denna uppställning är relevanta i det aktuella arbetet och finns implicit i programmerade regler. Exemplen är företrädesvis hämtade från Sköld (1966).

När det gäller nominalt led mellan finit verb och verbpartikel är det oftast så, som i Ex 42 a), att detta är subjektet men i vissa andra fall som b) *falla ... in,* är det uppenbart objekt.

Ex 42 a) I och med denna senare händelse *råkade SSF in* i vad vi kan benämna en krissituation. (jc14-094)
b) Det *föll mig in* att far och möjligen farfar också måste ha vanskött gården eftersom den så snabbt förvandlats. (kn08-019)

Ett ytterligare fall där subjekt tydligt kan identifieras som en icke-fundamenterad kandidat genom konstruktion är vid rektionsframflyttning och prepositionsstrandning som i Ex 43.

Ex 43 a) Resten av inkomsten lägger kommunen sedan beslag *på*. (eb09a-038)
b) Enda inslaget som inte direkt hade med Taube att göra svarade den käcka Ludvika spelmanstrio *för*. (cg03d-008)

Dessa fall finns med i den subjektsidentifikation som det aktuella programmet gör och som då kan ske på form och positionsmässiga grunder.[57] Subjektsidentifikation med ytterligare medel förutom position och struktur inklusive morfologisk taggning, finns i arbeten av t.ex. Øvrelid (2008). Där används en skala som inkluderar sådana formaspekter som morfologisk taggning men också animathet som måste listas för grupper av ord ofta harmonierar med syntaktisk funktionsprominens, dvs. benägenhet att fungera som subjekt hellre än objekt.

Ex 44 a) Animathetsskala: Mänsklig > Animat > Inanimat
b) Definithetsskala: Personligt pronomen > Egennamn > Definit NP > Indef NP
c) Syntaktisk funktionsskala: Subjekt > Objekt

I Ex 44 redogörs för tre skalor som tenderar att hänga samman och kunna ge rätt tolkning i framförallt transitiva huvudsatser, där rent konstruktionsmässiga ledtrådar saknas. Tolkningen av *Brevet (OBJ) skickade flickan (SUB)* blir t.ex. riktig genom iakttagelsen att *flickan* (mänsklig) finns högre än det inanimata *brevet* på skalan. Dessa arbeten beskriver också många undantag, exempelvis verbtyper som fungerar omvänt: *Flickan skrämde det höga stupet*. Skalornas betydelsemässiga/strukturella aspekter finns delvis med i den aktuella implementationen genom listningar av animata termer och verb med speciell funktion i det aktuella programmet, men det är inte ambitionen att göra en heltäckande listning av nominala huvudord som är animata, mänskliga, inanimata osv.

En inte alls ovanlig frågeställning är också vad som egentligen är subjekt respektive subjektiv predikativ i enkla konstruktioner med kopulaverb som *Den duktigaste var fru Svensson.* I den nyare Mamban (Teleman 1974) avgörs tolkningen av *Den duktigaste var fru Svensson* respektive *Fru Svensson var den duktigaste* helt konsekvent i sådana svåra fall genom att kalla fundamentledet för subjekt. I SAG ges heller inga entydiga svar: efter ett liknande exempel kon-

[57] Svårare typer som kan illustreras av exemplet i nedanstående citat där inga andra led i satsen enkelt kan hjälpa tolkningen rätt. Avgörandet av subjekt utgående från aspekt 3), betydelsen, har uppmärksammas under senare år. ”Om en som sitter i en stab får in meddelandet Fienden beskjuter egna trupper och tolkar det som det var de egna trupperna som besköt fienden, så skulle han med all rätt kunna anklagas för oförstånd i tjänsten. Möjligen skulle det vid straffutmätningen kunna tagas hänsyn till hans förflutna, om det framgick, att han en längre tid hade studerat nordiska språk” (Sköld 1966).

stateras: ”För det mesta är det i fall som dessa antingen pragmatiskt givet eller ovidkommande för tolkningen vilketdera ledet som är subjekt i satsen.” (Subjekt, § 32). Även om detta konstaterande accepteras kan tolkningen ju få betydelse när utvärderingar av korrekthet ska göras. I *SynTag* (Järborg 1986) verkar bedömningen vara annorlunda och snarare formgrundad: de understrukna leden är subjekt.

Ex 45 a) Det mest slående är den exklamatoriska, uppstyltade dialogen [...] (NFOART1045)
b) Det estetiskt mest tilltalande i denna bok var vissa bildsidor i färg [...] (NFOART0114)

Ex 45 verkar tyda på tolkningen att strukturen, den typiska predikativformen (adjektivfras), får avgöra tolkningen av detta led. Här valdes efter viss inspiration (och utan att ha sett SAGs och Mambans tolkningar) en tolkningsart som sammanfaller med *SynTag* här. Samma tolkning verkar göras i *Talbanken* (Einarsson 1976) i TIGER-XML-format[58] enligt Ex 46 (texten visas ej i detta format i exemplet här).

Ex 46 Det utmärkande för den moderna situationen är att mannen och kvinnan gemensamt tvingas att fråga: hur ska vi fördela våra yrkes- och hemuppgifter? (s4085)

Bestämning av antal förväntade nominala led i huvudsats och verbfras

I metoden leder det faktum att adverbial teoretiskt kan finnas i obegränsat antal per sats och att de kan ha både NP-form och andra former till att välja en strategi som först identifierar de led som det finns ett begränsat antal av (dvs. subjekt och objekt/predikativ), medan återstående led, även NP-formade, kan vara adverbial. Utgångsläget är att huvudsatser har subjekt och att samordnade finita verbfraser, som känns igen genom att de saknar fundament, saknar subjekt. Antalet objekt/predikativ för verb ges av valensdata men sätts även ofta manuellt. Det ska poängteras att en stor del av verben har en valens som innebär att flera olika antal nominala objekt/predikativ förväntas. Valensen innebär därmed ofta en kombination av möjligheter, t.ex. intransitiv och monotransitiv – noll eller ett objekt – förväntas.

Oftast gäller att antalet nominala kandidater ska vara lika med *subjektsantal* + *objekt/predikativantal*. Då kan subjekt och objekt väljas bland de aktuella kandidaterna. Emellertid tillkommer möjlighet för vissa led att fungera både som

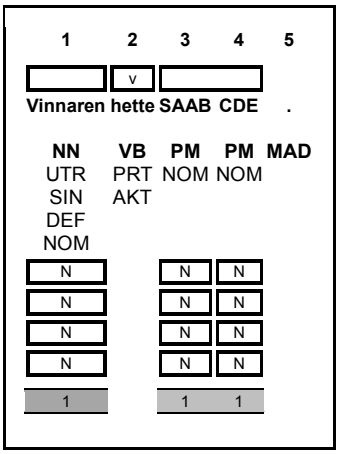
[58] Denna korpus finns att hämta på http://w3.msi.vxu.se/~nivre/research/talbanken.html (kontrollerad 11 september 2009).

ovanstående nominala led och adverbial, främst tidsadverbial. Antalet möjliga adverbial är, till skillnad från de nyss nämnda, inte begränsat. I många fall överstiger därför antalet kandidater det sammanlagda antalet subjekt, objekt och predikativ som förväntas. Satsen/verbfrasen i fråga påverkas därutöver av verbets modus och diates på följande sätt.

- *Verb i imperativ* leder till att inget subjekt förväntas.
- *Verb i passiv* (då ett objekt/predikativ generellt övergår till subjektsfunktion) innebär att valensdel med ett färre objekt/predikativ möjliggörs.
- *Konstruktion med formellt och egentligt objekt* leder till att det formella subjektet märks subjekt och det egentliga subjektet `objekt`.

Viss segmentförändring sker även efter skapandet av strukturnivå 3

Den översta strukturnivån, strukturnivå 3, är inte riktigt ekvivalent med den funktionella analysen utan utsatta syntaktiska funktionsetiketter. I steget där funktionella etiketter ska sättas undersöks antalet led så att exempelvis antalet nominala led i slutänden är lika med subjekt plus objekt/predikativled i en deklarativ huvudsats.



Figur 19 I ett fall som detta, där verbet (här verbet *hette*) generellt är monotransitivt och inget av de tre nominala strukturerna potentiellt är adverbial – såsom tidsadverbial – måste sammanfogning av två chunkar ske utan att någon regel för sammanfogning av dessa finns inskriven i programmet. Den enda möjligheten här är att sammanfoga *SAAB* och *CDE* för att få ett subjekt och ett objekt/predikativ, även om sådan explicit regel saknas. (Chunkar bestående av egennamn som varunamn inklusive modellbeteckningar är typiska kandidater för denna uteslutningsprocess.)

I den aktuella tekniken finns en fördel i att uteslutning av möjligheter används, jämfört med en grammatik som i princip måste ha en regel som matchar t.ex. en mycket ovanlig NP-struktur. Tyvärr ger inlemmandet av en sådan regel också en

risk för att denna kommer att användas vid felaktiga tillfällen – dvs. ju fler beskrivningsregler som används desto större blir risken för felaktiga strukturella analyser vid användning, vilket kanske ska ses som en av de absoluta svårigheterna med en grammatik som, till skillnad från här, innebär en explicit definition av språk. Ett fall som beskriver hur den heuristiska uteslutningsmetoden här fungerar även genom sammanfogning är en satsstruktur som Figur 19. Ett annat typfall av uteslutning är ju också hur fundament av godtyckliga längder sammanfogas till ett enda segment, nominalt eller adverbiellt, oberoende av struktur.

Utvärdering av primärsubjektsidentifikation

Subjektsidentifikationen är implementerad efter ovanstående beskrivning av segmentering och riktlinjer för identifikation, med många specialfall för identifikation i hjälpverbskonstruktion, anföringar etc. En uppskattning av det varierande antalet nominala ledkandidater enligt programmet redogörs för i Diagram 10.

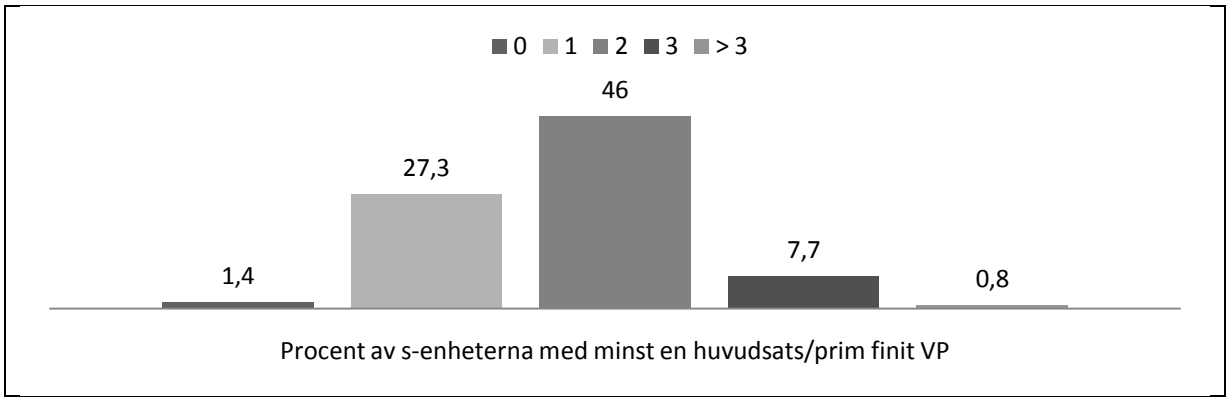


Diagram 10 En undersökning av antalet identifierade nominala led (dvs. kandidater för subjekt, objekt, predikativ och NP-adverbial) med programmet i första huvudsatsen i s-enheter med minst ett primärt finit verb genomfördes i omgångar på 1000 enheter och visade att det vanligaste antalet kandidater är två, men också att 27 % enbart innehöll en kandidat vilken oftast enkelt kan märkas rätt utgående från placeringskriterier: Subjekt krävs generellt, och blir därmed den etikett som väljs, men om det t.ex. är V1-konjunkt i imperativ är ledet objekt/predikativ.[59]

[59] Sökmönstret för denna undersökning är, exempelvis för sökning efter de s-enheter som har *en* kandidat:
`(mr.satser.length>0) && (mr.satser[0].FUNKKANDARR.length==1)`

858 av **1 000,** dvs. **85,8 %** slumpvis analyserade s-enheter var markerade av syntaxanalysatorn som innehållande *minst ett primärt subjekt*.

Frekvensuppskattning 12 Majoriteten av s-enheterna i ett test på 1 000 framslumpade s-enheter från träningsmängden innehöll markering av minst ett primärt subjekt.

En manuell undersökning av korrektheten för primärsubjektsidentifikation ger nedanstående resultat, enligt Utvärdering 1, för uppgiften att matcha *hela* subjektet, inklusive efterställda attribut, givet följande tolkningar:

- Det fanns en liten risk att samma enheter förekommer mer än en gång.
- Eventuella fel som antas bero på felaktig taggning i SUC 2.0 har ignorerats.
- I enheter med formellt och egentligt subjekt räknas det formella subjektet (i samtliga fall just här ordet *det*).
- I fall som *ni röker båda samma tobak* räknas det subjektskasusmarkerade ordet *ni* (ej ordet *båda*) som subjekt.
- I enlighet med Mambans resonemang ses *att*-sats/infinitivfras som subjekt i annars oklara fall, när denna konstituent skulle kunna utgöra egentligt subjekt i motsvarande sats med formellt subjekt.
- I eventuella verbellipser som *Kalle åt fisk, Lisa kött* räknas bara med ett subjekt per finit (här *Kalle*).
- Kommatecken inkluderat i slutet av en markering accepteras som rätt.

När det gäller klara regler för vad som är subjekt i kopulaverbkonstruktioner har här gjorts ett försök att göra de vanligaste tolkningarna. Om andra regler skulle bli aktuella är det dock generellt enkelt att ändra till dessa i programmet.

Helt korrekta markeringar (med rätt etikett)	Fel utsättning och/eller miss av det verkliga subjektet i huvudsats	Delvis matchning – ej fullständig täckning	För lång markering jämfört med det verkliga subjektet
375 av 400: **93,75 %**	17 av 400: **4,25 %**	5 av 400: **1,25 %**	3 av 400: **0,75 %**

Utvärdering 1 Korrekthet för precis identifikation av primära subjekt mättes manuellt. Utvärderingen gäller slumpvalda s-enheter med 400 huvudsatser ur testmängden, och är med reservation för den mänskliga faktorn. Felen exemplifieras i Tabell 28 och Tabell 29.

Felen i utvärderingen fördelade sig över väldigt många grupper, vilket kan ses som ett resultat av att konsekvent behandla grupperingar som svarar för många fel först i regelförbättringen mot träningsmängden. Felen verkade oftast bero på missade ordkategoriseringar och enkla misstag i regelformuleringar, generellt enkla att åtgärda, vilket ses som positivt, även om de fel som visas nedan inte har fått rättas till direkt, då de ju hör till testmängden.

Exempel på enheter som markerats som fel i testet	Felorsak
Redan i början av 20-talet kunde som nämnts ovan ***manskörens repertoar*** *betraktas som föråldrad.* (ja24-053)	*Som nämnts ovan* identifierades ej som eget adverbial utan är markerat i hela mittfältet. Subjekt ej utsatt alls.
Deltidsarbetande och lågavlönade *är systemets förlorare.* (ja15-053)	*Deltidsarbetande och lågavlönade* är ej listade som animata (mänskliga) – participfraser är normalt predikativ
Temperaturvariationerna *blir också allt mindre, ju större djupet är.* (fh08-126)	*Ju* har felaktigt setts som en del av en primär konjunktion och *är* är ej licensierat. Missad *som*-strykning.
Rätt öster om gården låg ***Epa-dalen med egna hem för arbetare på jord som utstyckats från Mölna Nedre****.* (kk01-098)	Fel huvudord och därmed chunktyp av *Rätt öster om gården* som är adverbial och ej NP
Hur högt, ja därom gick **meningarna** kraftigt isär i Ekonomiska klubben. (cg03a-017)	*Hur* har felaktigt licensierat *gick*

Tabell 28 Exempel på helt fel primärsubjektsmarkering från Utvärdering 1 visar många olika felorsaker utan att någon grupp direkt dominerar. Understrykningarna är parserns markeringar, det fetstilta segmentet är det önskade subjektet.

När det gäller resultatet för utvärderingen av exakt identifikation av primära subjekt men även de andra primära obegränsade leden så finns det en stor förbättringspotential med denna metod och det aktuella programmet. En stor del av felen antas vara enkla att åtgärda givet mer tid.

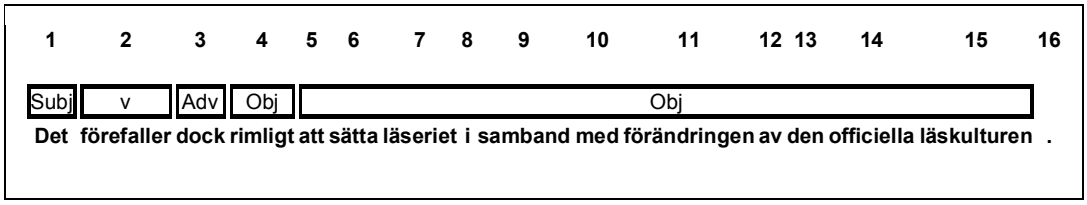
Exempel på enheter som markerats ofullständigt (bara delvis) samt de som markerats för långt i testet	Felorsak
[…] *och nu är* ***FFV:s tidigare*** ***generaldirektör Eric Malmberg*** *åtalad för smuggling via England till Saudi-Arabien.* (bb02a-029)	*Tidigare* har felaktigt blivit primärt adverbial
Något nytt och spännande kunde dock inte **dokumentären om utrikesdepartementet** bjuda på . (cg03g-013)	Regel för länkningen *dokumentären – om* saknades
I dag används **brytare med avancerade strömbegränsare och kontaktsystem**. (kk01-098)	Regel för länkningen *brytare – med* saknades
[…] och här kom **Kenneth Simonsson, SGIF**, in på 22:a plats. (ae06g-029)	Felaktig sammanfogning med efterföljande partikel/adverbial. (*in* borde dock vara taggat som partikel i SUC)
Mårten och Torsten - fast **de** brukar kalla honom Törsten. (kk26-025)	Missad förfältsidentifikation leder till för lång markering

Tabell 29 Exempel på ej fullständig identifikation (de tre första) samt för långa matchningar (de två sista) av primärt subjekt från Utvärdering 1. Understrykningarna är parserns markeringar, det fetstilta segmentet är det önskade subjektet.

Utvärdering av identifikation av objekt/predikativ/egentligt subjekt

Som redan klargjorts kallas alla de led som förekommer i N-positionen i satsschemats slutfält för objekt i kodningen, vilket även gäller för den anförda delen i en anföringssats. Däremot är syftet att hålla isär dem i segmenteringen, enligt Figur 20.[60] Som visas i Ex 47 räknas även s.k. *formellt objekt* som eget N-led. Arbetet med identifikationen av de primära obegränsade leden har främst fokuserat på subjektsidentifikation. Identifikation av N-led och adverbialled skulle kunna förbättras betydligt givet en längre tidsram.

[60] Det är en enkel lösning att kalla dessa objekt i XML-koden. Att här på vissa håll kalla dem N-led kan göra det svårt att typografiskt i kod skilja dem från n-led (subjekt).



Figur 20 Som nämnts benämns det egentliga subjektet i analysen, här en infinitivfras, för enkelhets skull *objekt* (jd01-051). Benämningen utgår från den placeringsmässiga likheten, N-positionen i satsschemat.

Ex 47 De ungdomar […] hade det ovanligt svårt att finna en bostad. (fc03a-006)

En utvärdering av identifikationen av N-led har genomförts med stränga bedömningskriterier, Utvärdering 2. I utvärderingen gäller att segmenteringen av N-ledet/N-leden ska vara helt riktig för att den ska räknas som korrekt, dvs. inklusive alla attribut. Detta formkriterium verkar därmed göra att denna utvärdering, som visserligen är manuell och utförd på korrekt taggade s-enheter från SUC, inte verkar kunna jämföras med andra parsningsansatsers resultat i litteraturen. S-enheter som inte antas kunna analyseras riktigt p.g.a. antagen feltaggning i SUC, eller p.g.a. diskontinuerliga konstituenter, har bortsetts från här. Felen visade sig i mycket hög grad gälla segmentering. Det innebar att objekt delats upp eller slagits samman med andra led, även med andra ledtyper som adverbial. I fallen med felaktig segmentering kan det dock nämnas att ledets huvudord oftast fanns med. Det kan också nämnas att felen oftast inte beror på den rangbaserade chunkningen.

---

622 av **1 000,** dvs. **62,2 %** slumpvis analyserade s-enheter var markerade av syntaxanalysatorn som innehållande *minst ett primärt N-led*.

Frekvensuppskattning 13 Majoriteten av s-enheterna i ett test på 1000 framslumpade s-enheter från träningsmängden innehöll markering av minst ett primärt N-led.[61]

[61] Sökmönstret i Javascript i programmet (se kapitel 4) var:
`(meningen.PRIMÄRLEDMÖNSTER_STR.indexOf('N') !=-1)`

Helt korrekta markeringar (med rätt etikett)	Fel segmentering men markering inklusive huvudordet	Fel etikett på markering	Övriga fel, beroende t.ex på licensiering
71 av 100: **71 %**	16 av 100: **16 %**	5 av 100: **5 %**	7 av 100: **7 %**

Utvärdering 2 En manuell utvärdering av identifikation av primära N-led, från s-enheter där sådana förekommer, med reservation för den mänskliga faktorn, visade att felen i hög grad består av felaktig segmentering, t.ex. att efterställda attribut ej markerats. Segmenteringsfel som lett till flera N-markeringar har bara setts som *ett* fel per huvudsats/primär finit VP.

3.5.2 Identifikation av primära obegränsade adverbial

Det följande avsnittet gäller primära adverbial inklusive både de begränsade varianterna, med många satsadverbial, och de obegränsade inklusive prepositionsobjekt, *som*-fraser och NP-formade adverbial. Adverbialstrukturer har i grundfallet inte potentiell nominal funktion, och identifikationen av dessa primära led är därmed inte sammanblandad på samma sätt som subjekt och N-led. Segmenteringsmässiga svårigheter återkommer dock även i identifikationen av adverbial. Inte nog med att adverbial kan finnas i teoretiskt sett godtyckligt antal per sats: de är placeringsmässigt fria på ett påtagligt sätt enligt satsschemat där de kan finnas i alla fält. De kan finnas både före och efter subjekt i mittfält, mellan infinitivmärke och infinit verb och som fokuserare före finit (se 3.6). Adverbial igenkänns dock nästan i alla fall (förutom de NP-formade) inte genom placering utan genom form. Rent allmänt är det enkelt att identifiera adverbial som ju oftast har PP- eller AdvP-form. Svårigheterna i denna uppgift, som innebär *precis matchning*, gäller i fall som när det är oklart om en sekvens av sådana strukturer ska sammanfogas eller inte.

En svårighet med *NP-formade* adverbial är att de kan fungera både som nominala och adverbiella led. En fråga som uppkom under arbetet var om det nödvändigtvis var så att en räkning av antalet nominala led krävdes så att rollerna subjekt och objekt/predikativ först måste tillsättas och om de övriga måste fungera på annat sätt. Det konstaterades tidigt att det fanns fall av ledtrådar av position och struktur i svenska som kunde avgöra tolkningen i hög utsträckning, statistiskt sett, utan att de olika leden måste räknas. NP med tidsnomen som huvudord i fundamentposition verkade rent kvantitativt nästan alltid fungera som adverbial (och kan direkt markeras som sådana) såvida de inte p.g.a. klara placeringsrestriktioner måste fungera som subjekt (*Denna sommar skulle bli lång*) eller som objekt i anföring (*Denna tid, sa hon*). I nedanstående inledande mindre undersökning ställdes frågan: I vilken kvantitativ utsträckning fungerar sådana potentiella adverbial i fundamentet verkligen som adverbial?

En sökning efter 100 enheter där fundamentet i första huvudsatsen upptogs av NP med tidsnomen (eller annat nomen som ger adverbialfunktion, vilket enbart gällde för en av enheterna, Ex 48 d) som huvudord, enligt den listning av sådana huvudord som programmet använder, se Appendix, visade att sådant fundamentinnehåll finns i ca 1 % av s-enheterna i SUC. Av NP-fundamenten i de 100 enheterna fungerade minst 90 som adverbial i denna position som i Ex 48. Inte alla fall blev korrekt analyserade av programmet vid detta tillfälle.

Ex 48
a) *Förra året* skickade Rymdbolaget upp sin hittills största raket, en Maxus. (ec18a-077)
b) *Kvällen därpå* stod han i dörren med ett stort fång röda rosor. (kk73-093)
c) *Ett tag* var ”Kattas” ledning cirka två minuter. (ae02c-008)
d) *Ett femtiotal meter norr om den gamla stenen och det skrämmande fyndet* hade hon funnit en rhododendronbuske som bjöd […] (kl18-086)

Av de få undantagen kan, förutom de ovan nämnda fallen, nämnas de konstruktioner som i Ex 49, vilka är tydliga exempel där även fundamentledet krävs som annat än adverbial för att besätta rollerna som verbvalensen kräver.

Ex 49
a) *Dagarna* var korta och trista. (kn04-013)
b) *Några magra år* är därmed att vänta. (ba03b-016)
c) *Tiden 25.43* för Jan är nytt rekord. (ae06g-004)
d) *1988* hade sin Åsa Domeij. (ec07b-005)
e) *Dagar med högtryck och lågtryck* ger ju också naturliga variationer […] (fh08-106)

Det är därmed inte helt givet att NP av detta slag i denna position är adverbial, men det är en stark tendens. Ett intressant faktum i sammanhanget är att fyra av de fem mest frekventa substantiven i svenska enligt Allén (1972) är sådana som gör NP till potentiellt adverbial när de är huvudord, 1: *år,* 2: *dag*, 3: *tid,* 4: *del*, 5: *gång* – närmast hämtat från Viberg (1990). Som synes finns i Ex 49 däremot inga potentiella NP-adverbial i fundamentet med objekt/predikativfunktion. En heuristisk regel markerar potentiella adverbial-NP i fundamentet som adverbial om subjekt redan är utsatt. Utvärdering 3 gäller korrektheten för alla typer av adverbial, samtidigt.

---

631 av **1 000,** dvs. **63,1 %** slumpvis analyserade s-enheter var markerade av syntaxanalysatorn som innehållande *minst ett primärt adverbial*.

Frekvensuppskattning 14 Nära två tredjedelar av s-enheterna i ett test på 1000 framslumpade s-enheter från träningsmängden innehöll markering av minst ett primärt adverbial.

Helt korrekta markeringar (med rätt etikett)	Fel segmentering men markering inklusive huvudordet	Fel etikett på markering	Övriga fel
77 av 100: **77 %**	9 av 100: **9 %**	3 av 100: **3 %**	11 av 100: **11 %**

Utvärdering 3 En manuell utvärdering av exakt identifikation av primära adverbialled, med reservation för den mänskliga faktorn, visade att de flesta markeringar är korrekta. Segmenteringsfel som lett till flera adverbial-markeringar har bara setts som *ett* fel per huvudsats/primär finit VP.

Sammanfattningsvis kan återigen påpekas att dessa korrekthetsresultat för de primära obegränsade leden skulle kunna förbättras, i många fall relativt enkelt, givet mer arbete med programmet.

3.6 Speciella textmeningstyper

Detta avsnitt beskriver en samling undantag och ovanligheter bland grammatiskt riktiga svenska satser. Det rör sig om satskonstruktioner som inte alltid enkelt passar in i en grundläggande positionsgrammatisk beskrivning, och om mer ovanliga ledtyper vars roller är mindre uppenbara. Satser med dessa fenomen analyseras korrekt i varierande grad, enligt en medveten avvägning som har beaktat deras frekvens och den arbetsinsats som krävs för täckning. Det grundläggande antagandet är dock att det mesta, om inte allting, kan täckas givet tillräckligt tid med denna manuella/halvautomatiska metod för förbättringen av syntaxanalysen. De följande fenomenen är någorlunda vanliga och innebär i flera fall att den gängse positionsgrammatiska analysen, så som den beskrivits, har behövts modifieras.

3.6.1 NA-rockader

Ett vanligt smärre undantag från satsschemat såsom det ursprungligen uttrycktes finns i möjligheten att låta adverbial föregå subjektet i mittfält eller efterdel. Att flytta adverbial till platsen före subjektet (det köpte jag *inte* ↔ det köpte *inte* jag) kan kallas *na-rockad* och huruvida det svenska satsschemat till och med borde formuleras som det norska nedan diskuteras bl.a. i Andréasson (2007). Den språkliga effekten av att låta subjektet komma efter adverbial i mittfältet som i Tabell 30, har beskrivits som ett annat sätt att framhäva det genom ren placering – förutom det som kan göras med fundamentering.

Forfelt		Midtfelt				Sluttfelt	
F	**v**	**a1**	**n**	**a2**	**V**	**N**	**A**
Derfor	*kan*	*ikkje*	*Noreg*	*enno*	*selje*	*bananar*	*till utlandet*
Idag	*har*	*nok*	*mange*		*sett*	*konkurransene*	*på TV*
Vi	*har*	*jo*			*sagt*	*det*	*mange ganger*
Dette	*må*	*vel*	*alla*		*klare*		

Tabell 30 Det norska huvudsatsschemat i *Norsk referansegrammatikk* (Faarlund, Lie och Vannebo 1997), s. 859 har positioner för adverbialled både före och efter subjektsplatsen i mittfältet. Adverbialet i a2-positionen är ofta ett satsadverbial. De tids-, sätts- och rumsadverbial (TSR-adverbial) som enkelt kan placeras i mittfält kallas ibland *ramadverbial* (*De skulle denna dag resa till Ystad*), medan s.k. valensadverbial som är semantiskt knutna till huvudverbet inte gärna står före detta verb *(*De skulle till Stockholm resa idag*) utom i fundamentet.

Denna variation täcks i implementationen utan att några speciella åtgärder krävs – dvs. eftersom metoden inte, till skillnad från ”en explicit formulerad grammatik”, innehåller en uttryckligen formulerad satsledsföljd, behöver denna möjlighet inte uttryckligen postuleras. Holm och Larsson (1980) tar upp pronominella objekt, placerade i mittfältet som en liknande avvikelse från satsschemats grundkonstellationer (*Känner du honom inte?; Han träffade dem aldrig*) eftersom satsadverbialen *inte/aldrig* anses finnas i mittfältet. Inte heller denna möjlighet behöver få speciellt stora konsekvenser för detta arbetes metod, då fält inte används så strikt för leddefinitioner. Ett annat fall av objekt i mittfältet är negerade objekt, se Tabell 47. Att placera långa adverbialled i mittfält gör ofta textmeningar stela eller formella: *De skulle i den mån vi andra kunde komma på besök möta upp*.

3.6.2 Fundamentdubbleringar

Dubbleringar av samma ledtyp förekommer på olika sätt i svenska. Dels förekommer ofta en dubblering av fundamentledet, ofta adverbialledet som typiskt dubbleras med adjunktionellt *så*, som i Ex 50. Ett adjunktionellt *så* påminner om ledfunktionen hos ett långt fundamentled, och kan här hjälpa till att identifiera V1-formad konditional. Fundamentdubbleringar kan även signalera t.ex. subjekt (t.ex. *han*) som i c) eller objekt (t.ex. *honom*) som i d). Ex 50 b) visar den stilmässiga effekt konstruktionen i övrigt får.

Ex 50
a) Vad det än var så var det tillräckligt bra för att få honom frikänd. (kn18-082)
b) Men hjärtat det var starkt. (ga05-096)
c) Och Torsten Bergman, han funderade. (kk21-071)
d) "Min man, honom skulle du ha dyrkat. (kk54-205)

Dessa fundamentdubbleringar, som i a) och b), analyseras i likhet med Mamban som del av det dubblerade ledet. Mambans beslut grundar sig främst på oviljan att placera flera led i fundamentposition, vilket är en relevant poäng även här. Ändå ska det kommas ihåg att det egentligen rör sig två led som inte gärna fungerar så vid parafras: *Vad det än var så var det tillräckligt bra för att få honom frikänd. ←//→ Det var tillräckligt bra för att få honom frikänd vad det än var Ø/*så.* När det gäller dubbelled med pronomen som avgränsats med komma som i c) och d) är tolkningen istället konsekvent att bara *han* respektive *honom* utgör fundamentled.

3.6.3 Verbanslutna fokuserare

Ett annat fenomen som utmanar idén att enbart ett led ska föregå finitet i huvudsats är fokuserare som i Ex 51.

Ex 51	a) Neologin inte bara betonade sedeläran […]	(jd01-107)
	b) Jag mer eller mindre föste ut honom.	(kk59-112)
	c) […] han till och med glömde det nya dataspelet.	(cg01b-015)

Under uppbyggnad har de adverbialled som ofta har denna ställning samlats in, men implementationen hanterar inte detta fenomen korrekt i skrivande stund.

3.6.4 *Kanske*-satser

Kanske-satser är ett omforskat grammatiskt fenomen som kan verka bryta mot satsschemat och V2. Svenska huvudsatser med det primära adverbialet *kanske* (enligt bl.a. Josefsson (2001), också *måhända/törhända*) kan parafraseras, med samma ord, på fler sätt än motsvarande satser med andra adverbial. Att detta beror på ordens etymologiska verbkaraktär är den förklaring som framförs bl.a. i Andréasson (2002) som sammanfattar svensk forskning om just *kanske*-satser.[62] Det kan först av allt konstateras att *kanske*, i likhet med ovanstående grupp (med *nästan*, *bara* m.fl.), ibland skulle kunna ses som verbmodifierande *fokuserare* i verbföregående ställning som i Ex 52 b), men det märkliga blir från den utgångspunkten exemplen i a) och c).

Ex 52	a) I någon mån kanske tanken låg i luften:	(fd02-095)
	b) Han kanske kan hjälpa er.	(kr05-096)
	c) Kanske det var därför han föreföll så tveksam?	(kn17-105)

[62] Det beskrivs intressant hur utvecklingen av *kanske* som ord med speciell syntaktisk funktion har skett i en process som försiggått parallellt också i andra europeiska språk.

Kanske-gruppens ord kan tillsammans med upp till två andra led föregå det primära finitet och därmed ge ”ett fundament med flera led”. Detta är för enkelhetens skull tolkningen som görs här, rent tekniskt, även om de positioner som föregår finitet inte behöver kallas fundament, eftersom detta kan vara missledande. För det praktiska syftet här är analysen helt enkelt att leden före finitet finns i ett fundamentliknande ”fält” med flera led, som i Tabell 31. Dessa led ses verkligen som primära och leder till att huvudsatser som innehåller ett ord från denna grupp får fler möjliga parafraser som antas kunna bildas med bibehållna sanningsvillkor. Eftersom *kanske*-orden alltså även *kan* verka verbmodifierande finns restriktioner i sanningsvillkorsbevarandet kopplat till somliga av omflyttningarna.

Förfält	Pos 1	*Kanske*-ord	Pos 2	Pos 3	Finit verb	Icke-finit verb	Objekt/ predikativ	Adverbial
Och	*idag*	*kanske*	*vi*	*faktiskt*	*skulle*	*välja*	*detta*	*ändå.*

Tabell 31 I en *kanske*-sats med finit verb kan sägas användas ett förenklat satsschema, ungefär som ovan, där subjekt ej har en kanonisk form i ”mittfältet” efter finitet. Subjektet finns däremot alltid på någon av positionerna före det verkliga finita verbet. För att identifiera subjekt används alltså en ansats som liknar den i hjälpverbkonstruktioner. (Vissa alternativa adverbialpositioner är möjliga, precis som ovan nämnda *na*-rockader i grundschemat).

Idén ”att *kanske* verkligen skulle vara ett finit verb”, vilken inte används här, kan sägas stödjas genom det faktum att huvudsatser kan konstrueras utan övrigt finit (*Idag kanske han kommit fram*) och att *kanske* kan föregås av *adjunktionellt så* när det föregås av adverbial, precis som primärt finit, (*Senare idag så kanske de ska spela*). Men i analogi med hur *bli*-konstruktionen (*bli hämtad*) här inte räknas som en hjälpverbskonstruktion, i satsschemats mening, räknas inte heller *kanske* i dessa satser som verb. Anledningen är i båda fall att ordklasstaggningen i SUC 2.0 inte ger komplementet (*hämtad* i exemplet) taggning som verb (utan som particip) eller ger *kanske* taggning som verb (utan som adverb).

Existensen av *kanske*-satser har istället föranlett en helt egen separat analysprocess för dessa i implementationen. Till skillnad från hur vanliga fundament generellt kan sammanfogas till ett enda led oavsett inre struktur innebär igenkänning av denna struktur att ett antal led, enligt Tabell 31, kan finnas före det finita verbet. Analysen av denna del inom vad som normalt kan kallas fundament påminner om analysen av andra fält. I fråga om subjektsidentifikation i denna konstruktion fungerar position 2 och 3 i Tabell 31 liknande ett mittfält i en hjälpverbskonstruktion: om NP finns på denna plats är det per default subjekt (med undantag för NP-formade adverbial), medan om det saknas kandidat där, så

finns det före *kanske*-ordet (dvs. i position 1 enligt Tabell 31). Subjektet måste alltså komma före det verkliga finita verbet. I konstruktion med egentligt subjekt gäller identifikationen det formella subjektet.

---

72 av **40 000,** dvs. **ca 0,2 %** slumpvis analyserade s-enheter var markerade av syntaxanalysatorn som innehållande *minst en kanske-sats*.

Frekvensuppskattning 15 Frekvensen av identifierade *kanske*-satser ligger på omkring 2 promille. Identifikationen är som nämnts inte helt fulländad.

3.6.5 Verbellips

När verbförekomster, finita eller infinita, underförstått upprepas utan att dessa ord finns med innebär det, som i Ex 53, att satser eller fraser startar utan att denna metod som utgår från verben ”söker efter nya satsled”.

Ex 53 a) Scenbilderna har varit konkreta och kostymerna realistiska. (gb02-115)
b) Bakfoten är 4-5 cm lång och framfoten 3-4 cm. (fh02-031)

Elliptiska uttryck där verbled utelämnas innebär en klar svårighet i systemet. Svårigheten ligger då ofta i det faktum att fler än de uppenbara huvudsatskonjunkterna finns närvarande – med andra ord är det fråga om satsvärdiga uttryck utan avgörande verb, och det är från början svårt att så att säga bevisa deras existens. Detta är möjligen en svårighet som gäller särskilt för denna verbfokuserande ansats med en satsnivå åt gången, och kanske även andra verbcentrerade ansatser, som dependensgrammatik. I den aktuella ansatsen byggs ju huvudsatser och primära finita verbfraser upp kring de primära finita verben. Ledtrådar till vad som försiggår finns ofta, som i exemplet i en speciell strukturform som två nominala led efter en (i en bemärkelse primär) konjunktion. För närvarande identifieras och hanteras dessa inte korrekt.

3.6.6 Finitsamordningar inklusive pseudosamordningar

Ex 54 a) Här studerar och forskar dessutom människor från ett hundratal […]. (hf01a-008)
b) Som sådan betraktades och förkunnades han av lärjungarna. (jd02-020)
c) Efter klippningen rensas och sorteras ullen i olika kvaliteter med tanke på [...] (fk01-044)

Pseudosamordningar (*Han sitter och läser*), vilka har uppmärksammats speciellt av Kvist Darnell (2008), och andra finitsamordningar, som i Ex 54, noteras särskilt. Anledningen till varför dessa bör uppmärksammas här för parsningen är att

analysen annars riskerar att identifiera övriga satsled fel. En pseudosamordning/finitsamordning har på ytan formen *”X – v – KONJ – v – X”,* där konjunktionen är primär och gäller i fallen här två primära finita verb. Det innebär att default-tolkningen blir att den första delen (fram till konjunktionen) skulle vara en huvudsats som slutar med finitet, medan den andra delen skulle vara en med den föregående satsen samordnad finit verbfras. Denna tolkning skulle leda till att de första delarna i före konjunktionerna i Ex 54 skulle sakna subjekt (det finns inget nominalt led som kandidat före verbsamordningen). Därför är det nödvändigt att särbehandla dessa s-enheter och schemamässigt välja tolkningen att finiten delar på v-positionen i vad som är en och samma huvudsats. Detta fenomen är inte helt korrekt hanterat i skrivande stund.

3.6.7 Diskontinuerliga konstituenter

Fenomenet diskontinuerliga konstituenter har dragit till sig en hel del intresse från forskningen som utgår från den generativa beskrivningsmodellen, se speciellt *Readings on Unbounded Dependencies in Scandinavian Languages* (Engdahl och Ejerhed 1982). Enligt denna modell är dessa former, inklusive utbrutna satsled och spetsställningar, hanterade med tomma element i de grundträd som blivit utsatta för olika transformationer. En rent ”linjär” beskrivning av grundläggande regler för acceptabel diskontinuitet i svenskan har också skisserats av Birn (1991). Den inleder med en grundläggande indelning i vänsterdiskontinuerlighet (*WH-flyttning* – i detta fall flyttning av *delar* av schemapositionsinnehåll) som i Ex 55 a) – c) och högerdiskontinuerlighet (*extraposition*) som i d).

Ex 55 a) Vem tänker du på just nu? (Diskontinuerlig PP, rektionsframflyttning)
b) Elvis borde nog alla ha en bild på. (Diskontinuerlig NP, med PP)
c) Svensson tror jag att (han sade att…) alla redan känner.
(Diskontinuerlig sats)
d) Jag träffar den minister i morgon som ansvarar för detta.
(Diskontinuerlig NP)

Typerna i Ex 55 har alla påträffats i SUC 2.0. En följd för satsanalyserna av exempel som i a) – c) är att subjektsidentifikationen där bör ta hänsyn till att del av objekt/adverbial spetsställts och välja nominalt led efter finitet som subjekt, vilket delvis sker. Exemplet i d) är däremot betydligt svårare och tolkningen i programmet är per default att relativsatsen är attribut till direkt föregående led, även om *i morgon* skulle kunna ingå i en listning av ”olämpliga strukturer för modifikation med relativsats”.

Förekomster av diskontinuerliga konstituenter upptäcks och hanteras inte fullt ut av föreliggande system som hittills fokuserar på huvudsatsled, och företrädesvis ouppdelade satsled. En regelrätt hantering av diskontinuerliga konstituenter gör

också att antalet möjliga satsmönster (och möjliga analyser per enhet) ökar. Samtidigt är det kanske rimligt att hanteringen av dessa skulle vara minst lika enkel med aktuellt linjärt förfarande jämfört med en trädstrukturell parsning. Eftersom utbrutna led innebär att ledmönstret, som i Ex 55, består av segment som är underordnade led eller delar av primära led uppkommer här en fråga om hur dessa segment ska etiketteras. Denna svårighet har naturligtvis även uppkommit i Mamban som gjorde en liknande linjär funktionell analys. När det gäller utbrytningar markerades utbrutet led + bisats i en utbrytning som *XX*, *obestämbar satsdel*, med exempel som (på s. 140): *Det (SUB) är (FIN) Svensson (XX) som han (SUB) pratar (FIN) med (PREP)*. Stroh-Wollin (2002) visar också den generella svårigheten som ibland finns när det gäller att avgöra om en sats innehåller utbrytning eller inte, enligt Ex 56. Versionen i b) försöker exemplifiera samma form utan vara en utbrytning (utan istället utgående från en grundsats som skulle kunna parafraseras *Lisa som lånade cykeln var det*).

Ex 56 a) – Det var Lisa som lånade cykeln. (Inte Anna.) – Utbrytning
b) (– Vem var det?) – Det var Lisa som lånade cykeln. – Ej utbrytning

3.6.8 Fria meningsled, satsinskott och apposition på satsnivå

Somliga segment i svenska textmeningar hamnar utanför den vanliga satsdelsanalysen. Det rör sig om ibland långa, ofta NP-formade, kommentarer eller specifikationer. Det finns andra typer av dubbleringar (utöver den ovannämnda fundamentdubbleringen) som nedanstående ’konstruktioner med paus’ hämtade från Thorell (1973), s. 271.

Ex 57 a) Tidningen, har den kommit än?
b) Nu har den kommit, tidningen som du frågade efter.
c) Tidningen har inte kommit, den heller.

Ex 57 a) analyseras helt enkelt genom att kalla *Tidningen* för förfältsled. Exemplen i b) och c) innebär dock vissa oklarheter. Från ett satsschematiskt perspektiv är det rimligt att utesluta *tidningen som du frågade efter* och *den heller* från den inre satsen, enligt Tabell 16 i kapitel 2. Satsinskott är ofta ett svårare fenomen att hantera, då de innehåller svårkategoriserade finit, som i Ex 58.

Ex 58 a) Mannen som hittade henne, Zawadzki Piotr, presenterade han sig militäriskt, stannade självmant i farstun […] (kk16-054)
b) Något som till en början roade henne […], men som - måste hon erkänna - i längden kunde bli en smula tjatigt. (kk03-069)
c) Kondoren, må ni tro, är en av de största rovfåglarna […] (kk30-049)

Apposition på satsnivå, som i *Det var den främsta kulturen, vilket var oomstritt* innebär en bisatsformad enhet som ses som adverbial med vissa placeringsre-

striktioner (eftersom det är tillbakasyftande). Igenkänning av apposition på satsnivå kan göras genom matchning av sekvensen ’kommatecken vilket’ som ej föregås av substantiv i neutrum. Detta sker inte helt korrekt i implementationen just nu.

3.6.9 Flerordstitlar

Titlar på böcker, film, musik etc. utgör en allmän svårighet för syntaktisk parsning, speciellt när dessa innehåller finita verb och rent strukturellt utgör satser. I svenska där *book title capitalization*[63] oftast inte används, skulle en identifikation av dessa behöva analysera segmentet som följer på ett ord med stor begynnelsebokstav fullständigt för att identifiera och sammanfoga hela namnet. På ungefär motsvarande grunder är texter där ord och uttryck behandlas på ett språkvetenskapligt sätt också svårhanterligt.

Ex 59 a) Igår Povel Ramel med Var är tvålen, broder. (cg03b-009)
b) Tjeckerna ska få se Dario Fos "Inget går upp mot mammas gräs" och ungdomspjäsen "Rymmarna". (af01o-006)
c) Dikterna av Pär Lagerkvist var bl.a. Vem gick förbi min barndoms fönster, Tacka vill jag blommor och molnen och Tillfällig som en vallmo. (af07o-005)
d) Vilken relation det handlar om ingår inte i betydelsen av ordet "är", utan detta bestäms av betydelserna av de andra orden i satsen. (fa05-099)

I det sista av exemplen handlar själva texten om ordet *är,* vilket är ordklasstaggat som finit verb och eftersom SUC innehåller en del språkvetenskapliga texter blir detta en egen kategori undantag. Även om viss heuristik kan specialbehandla de segment som står inom citationstecken är täckningen här ofullständig. Det är tveksamt om något system skulle kunna hantera uppgiften väl, även om den kan ses som en del i det aktuella forskningsfältet *named entity recognition (NER)*.

3.6.10 Skriftrelaterade svårigheter

I arbeten med text kan en del teckenhantering riskera att bli svårhanterlig och innebära tidskrävande arbete. När det gäller citattecken och parenteser är dessa i SUC 2.0 och i den tokenisering som görs vid hantering av annan text urskilda som egna enheter, även om de i text angränsar ord. Uttryck inom parentes undantas från början – verb etc. tillåts alltså inte samverka syntaktiskt med utanförvarande uttryck så att det utanförvarande har beroende till det inom parentes (antagandet är alltså att borttagandet av parentes med innehåll alltid ska efter-

[63] *Book title capitalization* innebär oftast att alla ord som ej är funktionsord får stor begynnelsebokstav, t.ex. *The Lord of the Rings*.

lämna en grammatiskt korrekt struktur). Undantag gäller för de s-enheter som helt finns inom ett parentesuttryck. Inskott av citattecken (*”* samt *’*) i textmeningar kan likaledes potentiellt hindra sammanfogning mellan två segment eller matchning av flerordsuttryck. För att hindra detta tas dessa tecken inledningsvis bort från textmeningar för att återkomma i utdataversionen.

3.7 Resultatens relation till andra moderna system för parsning av svenska

Som nämnts är det analysformat som produceras i schemaparsning annorlunda jämfört med det som produceras av andra system. För att ändå kunna relatera den aktuella metodens resultat och korrekthet hittills till andra system kommer här, delvis på begäran, några relevanta resultat att visas. Tre moderna systemtyper för parsning av svensk text är parsning med vad som på svenska har benämnts *restriktionsgrammatik* (*Swedish Constraint Grammar, SweCG*) (Karlsson, o.a. 1995)/(Voutilainen 2001) och parsning till en dependensgrammatisk analys med antingen *induktiv dependensgrammatikparsning* med *MaltParser* (Nivre, o.a. 2007) eller *Swedish Functional Dependency Grammar*[64] (*SweFDG, Conexor*) (Voutilainen 2001). De system som här nämns har grammatiska funktioner som del av analysen (eventuellt valbar i *SweCG*) och är i alla fall inte enbart helt frasstrukturella. Detta är anledningen till att de över huvud taget tas upp i sammanhanget, som påpekas återigen längre ner är formatet och förutsättningarna egentligen alltför olika för tydliga resultatjämförelser.

```
1 Den          det:>3
2 gamla        attr:>3
3 hunden       subj:>8
4 och          cc:>3
5 den          det:>7
6 unga         attr:>7
7 katten       cc:>3
8 äter         main:>0
9 frukost      obj:>8
```

Kodexempel 6 I *SweFDG* (Voutilainen 2001) produceras dependensgrammatisk analys, här i textformat. Exemplet är förkortat och information om ordklasstaggning och ordens grundformer är borttagna för enkelhets skull. Den version av SweCG som utvärderats i samma artikel resulterar i en liknande funktionell analys. Analysen i textformat visar hur länken från huvudverbet *äter* går till subjektets huvudord *hunden* (bågen från *hunden* pekar på *äter*).

Constraint Grammar (*CG*) är, som nämnts, ett ramverk som syftar till att enkelt formulera restriktioner i regelval när en syntaxanalys sker, och det riktar därmed in sig på den stora ambiguitet som en normal grammatik kan ge upphov till. *CG*

[64] Programmet *FDG* distribueras liksom *Constraint Grammar* i skrivande stund genom finska *IT Center for Science*: http://www.csc.fi
MaltParser finns tillgänglig på http://w3.msi.vxu.se/~nivre/research/MaltParser.html (webbplatserna besöktes 20090719).

tillåter regelskrivning både för ordklasstaggning, och olika typer av syntaxanalys, bland annat en som är jämförbar med *SweFDG*, se Kodexempel 6. Ramverket *CG* är inte fritt tillgängligt utan har blivit avgiftsbelagt.

	Subjekt (huvudord)	Objekt (huvudord)	Predikativ (huvudord)
Precision	98 %	95 %	97 %
Recall	92 %	90 %	95 %

Tabell 32 *The Swedish FDG parser* utvärderades genom en undersökning av hur väl subjekts, objekts och predikativs huvudord kunde länkas till deras regenter – dvs. huvudverben. Undersökningen gjordes manuellt på 406 textmeningar från *Hufvudstadsbladet* och *Dagens Nyheter*. Detta system får betraktas som ett av de som har den bästa korrektheten för svenska idag. Det är emellertid svårt att jämföra denna prestanda med andra system.

	Subjekt (huvudord)	Objekt (huvudord)	Predikativ (huvudord)
Precision	95 %	94 %	92 %
Recall	83 %	88 %	96 %

Tabell 33 Tapanainens and Järvinens resultat med SweCG enligt Voutilainen (2001).

I Wilhelmsson (2008) redovisades en manuellt beräknad sammanvägd korrekthet (*F-score*) på ca 98 % för identifikation av primära finita verb i korrekt taggad text från en testdel av SUC 2.0. I de andra systemen är denna korrekthet lägre, när den alls nämns (testtext och ordklasstaggning är dock olika och gör det i princip omöjligt att jämföra rättvist). De tre nämnda systemens korrekthet ifråga om identifikation av *huvudord i funktionella led* som subjekt har utvärderats och försök att jämföra dessa har gjorts av upphovsmännen. Som poängteras i Citat 3 är det till och med svårt att jämföra så relativt närliggande system som *MaltParser* och *SweFDG.* Redovisade resultat ges i Tabell 32 till Tabell 34.

	Subjekt (huvudord)	Objekt (huvudord)	Predikativ (huvudord)
Precision	90,6 %	78,9 %	75,9 %
Recall	88,1 %	83,5 %	71,8 %

Tabell 34 Resultat från en betydligt noggrann utvärdering mot *Talbanken* (Einarsson 1976). ”We use the professional prose section, consisting of material taken from textbooks, newspapers and information brochures” (Nivre, Hall, o.a. 2007), s. 113. Undersökningen gjordes med en ordklasstaggning av *Talbanken* med en korrekthet av 95,6 % i utvärderingen av *MaltParser*.

Citat 3 *Relating the Swedish results to the state of the art is rather difficult, since there is no comparable evaluation reported in the literature, let alone based on the same data. Voutilainen (2001) presents a partial and informal evaluation of a*

> *Swedish FDG parser, based on manually checked parses of about 400 sentences from news paper text, and reports F measures of 95 % for subjects and 92 % for objects. These results clearly indicate a higher level of accuracy than that attained in the experiments reported here, but without knowing the details of the data selection and evaluation procedure it is very difficult to draw any precise conclusions.* (Nivre, Hall, o.a. 2007).

Eftersom resultaten rörande *SweFDG* (Voutilainen 2001) kommer från ett test på en relativt ospecificerad text av oklar svårighetsgrad har de ansetts för oklara för att ens jämföra med liknande resultat från den mest närliggande systemtypen – den (likaledes funktionella) dependensparsern *MaltParser* enligt Citat 3. Det är uppenbart mycket stora skillnader emellan undersökningarnas förutsättningar och den output de producerar, varför en jämförelse riskerar att bli helt missledande. Ännu svårare är det rimligen att jämföra resultaten med föreliggande system. *The Swedish FDG parser* har testats på text som den ordklasstaggat själv. I det föreliggande arbetet är det istället syntaxanalysen och inte taggningen som varit fokus. Taggaren som använts där har inte någon speciell korrekthet angiven i källan. *The Swedish FDG parser* har i likhet med *MaltParser* utvärderats med identifikation av funktionella länkar både på huvud- och bisatsnivåer.

I fråga om korrekt länkning från satsers huvudverb till subjektets huvudord (på både huvudsatsnivå och underordnade nivåer) redovisas, som omtalas i Citat 3, för *SweFDG* en F-score på ca 95 % i ett manuell test, jämfört med motsvarande föredömligt objektiva utvärdering av resultatet från de inducerade reglerna i *MaltParser* till ca 89,3 %. I *MaltParser* används en ordklasstaggning med en korrekthet på 95,6 %, i *SweFDG* är motsvarande värde okänt. Till *MaltParsers* fördel ska dessutom konstateras att det är den använda korpusen som avgjort vad som räknas som subjekt, de oklarheter och skillnader angående vad som räknas skulle t.ex. kunna betyda att den egentligen har samma korrekthet som *SweFDG*.

Att relatera dessa olika värden till de som hittills uppnåtts med aktuell metod är rimligen ännu svårare. Utvärderingen ger över 90 % korrekthet för matchning av *hela* huvudsatssubjekt i korrekt taggad text (testmängden från SUC 2.0). Det ska då beaktas, till detta projekts fördel, att 1) en utökad analys till underordnade satsnivåer kanske skulle ge högre korrekthet p.g.a. subjektets regelmässiga placering där, samt 2) att ytterligare höjning troligen skulle ske om identifikation av huvudord vore tillräckligt. Dessutom har s-enheter med antagen feltaggning undantagits från korpusen när sådana påträffats här, så taggningen har inte tillåtits leda till fel. Fastän fullödiga jämförelser därmed svårligen låter sig göras så har föreliggande system en stor potential till utveckling. Metoden för parsning är specialiserad för svenska och har en väldigt rik intern representation av den textmening som analyseras (se Kapitel 4). Det ger möjlighet till smidiga ändringar och komplex regelskrivning som tar t.ex. valens och pågående delanalys i beaktande mer än för närvarande.

De korrekthetsvärden som visas upp är resultatet av en lång periods manuellt arbete, i hög utsträckning genom stegvis felrättning. Fel har sorterats i kategorier vilka genomgås med förhoppningen att nå fram till allmängiltiga, handskrivna analysreger. Detta betyder att resultaten inte speglar kvaliteten hos metoden som sådan, utan bara det aktuella programmets status. Eftersom regelskrivandet i den aktuella implementationen sker mycket fritt och kan uttrycka i princip alla grammatikregler som grundar sig på form och ordning, verkar det möjligt att nå mycket långt – att utvecklingen tar tid är mer av en begränsning. Resultaten som erhållits med denna metod genom implementationen är inte riktigt möjliga att jämföra med något av de nämnda systemen, åtminstone beroende på följande anledningar.

- Programmet har som målsättning att identifiera *hela* satsled exakt från första ordet i första framförställda attributet till det sista ordet i det sista efterställda attributet.

- Det föreliggande systemet har konskevent utvecklats mot korrekt taggning, eller enheter med ofarliga taggningsfel från SUC 2.0. Tillämpningarna i kapitel 5 använder sig av en taggare med uppskattad korrekthet 95,3. Andra system är utvärderade mot ordklasstaggning från ca 95 procents korrekthet.

- I projekten *SweFDG* och *MaltParser* läggs utvärderingsfokus naturligt på korrekt identifikation av *huvudord* hos subjekt och objekt.

- Programmet är i sitt nuvarande utförande kapabelt att göra huvudsatsanalys men inte fullständig satslösning av svenska textmeningar. Även om mycket talar för att analys av satsförkortningar, bisatser och relativsatser kan analyseras lika väl med samma metodik (ordföljden verkar till och med göra uppgiften enklare) så är detta ännu ej utarbetat. Programmet i sin nuvarande form har speciellt svårigheter med vissa elliptiska konstruktioner, *Han köpte mjölk och hon smör* – eftersom *hon smör* motsvarar en huvudsats, men saknar finit och ansatser bygger huvudsatser och primära finita verbfraser kring finit.

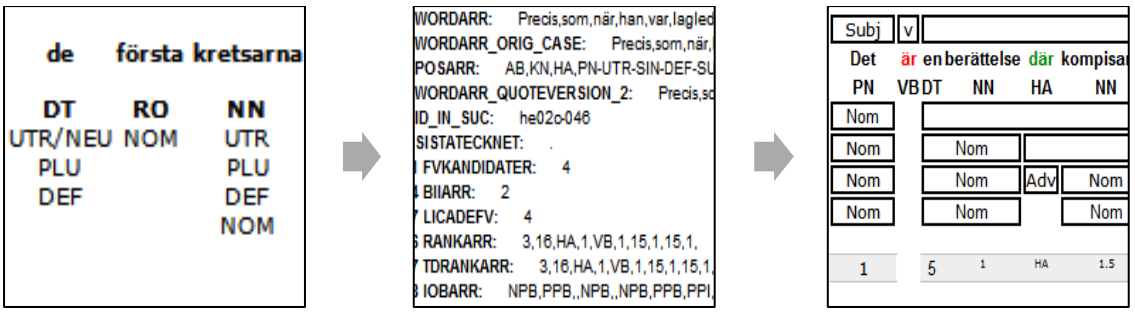
- Jämfört med *GTA* (Knutsson, Bigert och Kann 2003), men troligen även jämfört med *MaltParser*, är uppmärkningen av SUC omodifierad och innehåller inte från början tillagd information som speciell taggning av hjälpverb, kopulaverb etc. Däremot finns i systemet många ordgrupperingar varmed ord märks upp under syntaxanalysen.

Sammanfattningsvis i detta kapitel om identifikation av de primära obegränsade leden kan sägas att ledidentifikationen delvis har visat höga korrekthetsresultat. Speciellt beräknas identifikationen av primära subjekt kunna ske med hög kor-

rekthet. Detta är i hög grad en konsekvens av att denna uppgift ägnats mycket tid. Utvärderingen av de obegränsade leden har generellt gällt exakt matchning inklusive alla attribut. Felmarkeringar är hittills ofta felaktiga segmenteringar p.g.a. att otillräckligt med tid ägnats åt olika länkningsregler. Den rangbaserade chunkningen är däremot inte någon kvantitativt betydande felkälla.

4 Tekniskt utförande

Detta kapitel svarar på frågan hur de relativt allmänt uttryckta reglerna i de föregående kapitlen kan implementeras i en fri formalism som inte följer de regelformat som allmänt förekommer vid parsning enligt uppställningen i kapitel 1, vilka ofta innehåller en bidirektionell grammatik som även fungerar för generering. Den representationsstruktur som istället används för textmeningar under analys, enligt Figur 21, är mycket informationsrik och enkelt utbyggbar med ytterligare informationssärdrag och klassificering av ord i grupper som fungerar på särskilda sätt syntaktiskt. Detta har lett till en positiv grundsyn i fråga om systemets potential. Väldigt få textmeningar är nämligen så komplicerade i detta perspektiv att den information som behövs för korrekt analys inte ska kunna uttryckas på något sätt med nämnda medel som handlar om form (inklusive ordlistningar), struktur och position.



Taggad indata från SUC 2.0 kommer (otaggad textinput ordklasstaggas med HMM-taggare)

Analyssteg 1

Insamling av egenskaper, licensiering av begränsade led, och rangbaserad chunkning i objektet ***Mening***

Analyssteg 2

Fortsatt analys och identifikation av obegränsade led, uppbyggnad av chunkstruktur, främst med objektet ***MR***

Figur 21 Under analysen samlas grundläggande fakta om indata, licensiering och grundläggande chunkning företrädesvis i objektet *Mening* (Analyssteg 1). I den följande huvudsatsvisa, och primära finita VP-visa, analysen placeras representationen av den mer avancerade sammanfogningen av segment och identifikation av obegränsade led huvudsakligen i objektet *MR* (Analyssteg 2). Dessa två datastrukturer som beskrivs i avsnitt 4.1 lagrar både indata, delanalys och slutresultat. Figur 28 ger en betydligt mer detaljerad bild av förloppet, tekniskt sett.

Struktur i kapitel 4, Tekniskt utförande

Detta kapitel består av relativt fristående delar som beskriver representationsformat, ingående delfunktionaliteter, gränssnitt och arbete med valensextraktion från lexikon.

Avsnitt 4.1 behandlar de övergripande datastrukturer som används internt för representation av analysen såsom den enligt denna metod sker i delsteg. Eftersom analysen inte sker med en frasstrukturell fullständig trädstruktur eller enligt någon annan typisk intern datarepresentationsmodell används här istället två stora datastrukturer som innehåller både direkt härledbar information, delresultat och till slut resultatet för analysen av en textmening. Processen som detta kapitel beskriver behöver inte ses som en del av själva metoden som sådan. Det är fullt möjligt att representera syntaxanalysen på annat sätt.

Avsnitt 4.2 beskriver det praktiska arbetet med analysförbättring som sker under användning av det aktuella programmet, företrädesvis mot SUC 2.0. Arbetet innebär huvudsakligen insamling av felaktiga analyser och kategorisering av dessa för att möjliggöra systematisk manuell förbättring, genom regler som beskrivs med de två datastrukturerna. Den tid som ägnas åt denna uppgift ger direkt tydlig förbättring i resultatet och det praktiska arbetet har fått en karaktär av ”buggfixning”.

Avsnitt 4.3 tar upp en ordklasstaggare som byggts för att tillåta arbete med fri textinput. Denna ordklasstaggning görs med en trigrambaserad dold Markovmodell (*HMM*).

Avsnitt 4.4 är en beskrivning av gränssnittets delar. Dessa har en betydelsefull roll i denna manuella ansats för att förbättra analysen, t.ex. genom att erbjuda ett sätt att snabbt kategorisera ord som har speciellt syntaktiskt funktionssätt.

Avsnitt 4.5 redogör för den användning av valensdata från databaser till uppslagsverk som görs i implementationen. Utgående från två stora lexikon utvinns somlig information framförallt för användning för PP-attachment frågor, dvs. för att sammanfoga efterställda attribut till föregående segment. För att använda dessa valenser har också en särskild uppslagsfunktionalitet skapats.

4.1 Utveckling med en objektbaserad representation av textmeningar

Detta avsnitt beskriver den tekniska representationen av de analyserade enheterna. Dessa har när de kommer från SUC redan ordklasstaggning, och i annat fall

märks de upp med ordklasstaggning med den funktionalitet som beskrivs i avsnitt 4.3.

De regler som i föregående kapitel beskrivits och utvärderats har antagit en mycket annorlunda form jämfört med en grammatik som bygger på omskrivningsregler, t.ex. en kontextfri grammatik. Eftersom metoden som här beskrivs inte begagnar sig av en formell grammatikformalism som de i Chomskyhierarkin, avknoppningar därifrån, *HPSG* (Pollard och Sag 1994) eller *LFG* (Bresnan 2001), utan från en traditionell modell som inte implementerats fullskaligt för fri text tidigare, så har inre representation och presentationsformat för resultatet blivit en kärnfråga under arbetets gång. För den inre representationen av en aktuell textmening finns två övergripande datastrukturer (*objekt*) varav den ena är en löst sammansatt samling av information som till stor del är uttrycklig och enkelt härledbar, medan den andra är ett mer noggrant utformat objekt som inbegriper andra objektstrukturer och som speglar struktur och funktion och vari den slutliga grammatiskt funktionella representationen placeras före generering av resultatformat. Resultatet finns i ett XML-format och visuellt med HTML. Denna representation har visat sig användbar och ordnar upp de variabler som sätts under analys, även om det inte hävdas att den skulle vara optimal eller slutgiltig.

Objekten som representerar aktuell textmening, *mening* och *MR*, innehåller direkt härledd indata, som positioner för verb, längd etc., men också analysens del- och slutresultat, som positioner för licensierade och primära verb, huvudsatser och primära satsled.

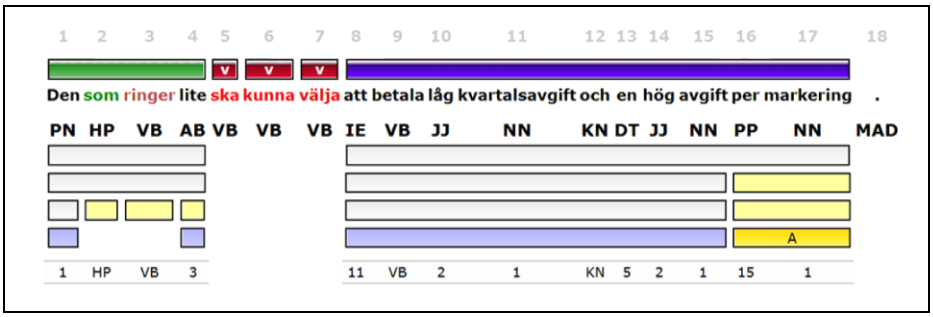
4.1.1 Objektet *Mening*

När en textmening inledningsvis undersöks placeras grundläggande data i en attribut-värdematris, implementerad som en objekt-datastruktur med en enda nivå, *mening*. I denna placeras både direkt information såsom längd, ordpositioner och ID från SUC samt andra värden som härleds från de första genom mekanismerna för identifikation av begränsade primära led och licensieringar. Objektstrukturen *mening*[65] har ett hundratal ingående möjliga attribut-värdepar, där viss redundans förekommer och alla värden inte instantieras för alla s-enheter. Dessa värden beskriver generella egenskaper för hela meningen och index för olika konstituenter, enligt Tabell 35.

[65] I grammatiklitteraturen är begreppet *mening,* som redan nämnts, flertydigt och till synes omtvistat. Denna diskussion lämnas därhän eftersom datastrukturerna här som i resten av arbetet gäller *s-enheter* såsom dessa framträder i SUC och i texter som analyseras, även om dessa inte alltid, eller enligt alla synsätt, borde kallas meningar.

Namn	Datastruktur	Beskrivning
LEN	Heltal	Längd (där interpunktioner etc. räknas precis som ord)
WORDARR	Vektor (endimensionellt fält)	Ingående ord, t.ex. [”Det”, ”fungerar”, ”bra”, ”.”])
POSARR	Vektor (endimensionellt fält)	Ingående ordklasstaggar
BIIARR	Vektor (endimensionellt fält)	Index (positioner) för tydliga bisatsinledare
FVKANDIDATER	Vektor (endimensionellt fält)	Index (positioner) för tydliga finita verb
IVKANDIDATER	Vektor (endimensionellt fält)	Index (positioner) för tydliga icke-finita verb
FUNDAMENTS	Vektor (endimensionellt fält)	Vektor (parvis start- och slutindex per fundament): [0,0, 7,12]. Ett par där första elementet är högre än andra betyder ”fundamentlös”. Paret kan t.ex. vara 5,4 – då är konjunktens pfv på position 5.
IS_FRAGA	Boolean	*True* om meningen antas vara en fråga
LICADE_FV	Vektor (endimensionellt fält)	Vektor med index för licensierade finit
PPSTARTS	Vektor (endimensionellt fält)	Index för prepositionsfrasstarter
UNB_FALT	Vektor (endimensionellt fält)	Index för start- respektive slutindex parvis för de segment som upptas av obegränsade konstituenter
ID_IN_SUC	Sträng	Textmeningens eventuella ID i SUC, t.ex. ’kn03-182’

Tabell 35 Exempel på några centrala variabler i objektet *mening*. Vektorer (’arrayer’) som används är *nollbaserade (zero based)*, dvs. första elementet har index 0, det andra 1 osv. För behandling av de begränsade leden som finita verb, partiklar osv. gäller att de har en struktur för insamlade kandidater, en för licensierade instanser och en för primära instanser (index för ordpositionerna i s-enheterna).



Figur 22 Nedanstående Kodexempel 7 svarar mot denna s-enhet (af03l-002).

Kodexempel 7 exemplifierar värden som initieras och sätts under analys i objektet `Mening` för en vanlig s-enhet: den i Figur 22.

1 WORDARR_ORIG_CASE_INCL_QUOTES:
Den,som,ringer,lite,ska,kunna,välja,att,betala,låg,kvartalsavgift,och,en,hög,avgift,per,markering,.
2 POSARR_INCL_QUOTES: PN-UTR-SIN-DEF-SUB/OBJ,HP---,VB-PRS-AKT,AB-POS,VB-PRS-AKT,VB-INF-AKT,VB-INF-AKT,IE,VB-INF-AKT,JJ-POS-UTR-SIN-IND-NOM,NN-UTR-SIN-IND-NOM,KN,DT-UTR-SIN-IND,JJ-POS-UTR-SIN-IND-NOM,NN-UTR-SIN-IND-NOM,PP,NN-UTR-SIN-IND-NOM,MAD
3 WORDARR: Den,som,ringer,lite,ska,kunna,välja,att,betala,låg,kvartalsavgift,och,en,hög,avgift,per,markering,.
4 WORDARR_ORIG_CASE:
Den,som,ringer,lite,ska,kunna,välja,att,betala,låg,kvartalsavgift,och,en,hög,avgift,per,markering,.
5 POSARR: PN-UTR-SIN-DEF-SUB/OBJ,HP---,VB-PRS-AKT,AB-POS,VB-PRS-AKT,VB-INF-AKT,VB-INF-AKT,IE,VB-INF-AKT,JJ-POS-UTR-SIN-IND-NOM,NN-UTR-SIN-IND-NOM,KN,DT-UTR-SIN-IND,JJ-POS-UTR-SIN-IND-NOM,NN-UTR-SIN-IND-NOM,PP,NN-UTR-SIN-IND-NOM,MAD
6 WORDARR_QUOTEVERSION_2:
Den,som,ringer,lite,ska,kunna,välja,att,betala,låg,kvartalsavgift,och,en,hög,avgift,per,markering,.
7 ID_IN_SUC: af03l-002
8 SISTATECKNET: .
9 RAPPFVARR:
10 FVAUXARR: 4
11 FVKANDIDATER: 2,4
12 INFMARKARR: 7
13 IVKANDIDATER: 5,6,8
14 BIIARR: 1
15 BIIARR2:
16 SATSADVKANDIDATER:
17 LICADEFV: 2
18 LICADEIV:
19 LICADESATSADV:
20 FUNDAMENTWORDARR: Den,som,ringer,lite
21 FUNDAMENTPOSARR: PN-UTR-SIN-DEF-SUB/OBJ,HP---,VB-PRS-AKT,AB-POS
22 MITTFALTLEDARR:
23 FUNDAMENTLEDARR:
24 EFTERFALTLEDARR:
25 FALTARR:
26 RANKARR: 1,HP,VB,3,,,,11,VB,2,1,KN,5,2,1,15,1,
27 TDRANKARR: 1,HP,VB,3,,,,11,VB,2,1,KN,5,2,1,15,1,
28 IOBARR: NPB,,,NPB,,,,NPB,NPI,NPI,NPI,NPI,NPI,NPI,NPI,PPB,PPI,
29 PRIM_IV: 5,5,6
30 PRIM_AUX_FV: 4
31 PRIM_FV: 4
32 PRIM_KONJ:
33 FORFALT:
34 PRIM_REFL:
35 PRIM_PL:
36 PRIM_SATSADV:
37 LICADE_REFL:
38 LICADE_PL:
39 LICADE_SATSADV:
40 MAIN_CLAUSES: 0,17
41 FORNAMNSARR:
42 EFTERNAMNSARR:
43 PRIM_SLUT: 17

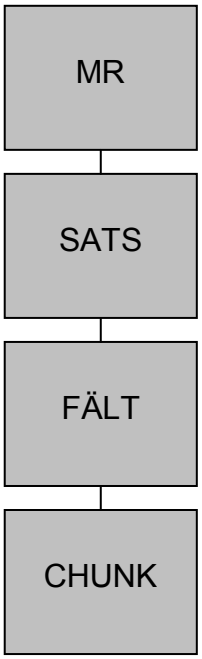
```
44 FUNDAMENTS:     0,3
45 UNB_FALT:     0,3,7,16
46 LEN:     18
47 LICADEFVAUX:
48 TDIOBARR:     NPB,,,NPB,,,,NPB,NPI,NPI,NPI,NPI,NPI,NPI,NPI,PPB,PPI,
49 MITTFALT:
50 SUBJEKT:
51 is_PASSIV:
52 NPARR:     0,1,3,1,7,8
53 PPARR:     15,2
54 NPHEADARR:     0,,,3,,,,14
55 PPHEADARR:     ,,,,,,,,,,,,,,,15
56 NPSTARTS:     0,,3,,7
57 PPSTARTS:     15
58 NPLENGTHS:     1,,1,,8
59 PPLENGTHS:     2
60 is_JANEJFRAGA:     false
61 is_FRAGA:     false
62 is_HVFRAGA:     false
63 EFTERLEDSSTRUKTUR:
64 FINKONJSTRUKT:     0,17,v2,HJV
65 k2arr:
66 STACKDJUPARR:     0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1
67 PARENTESINDEXOBJ:     [object Object]
68 ANVÄNDER_KONDLIC:     false
69 HELA_STRÅNGEN:     Den som ringer lite ska kunna välja att betala låg kvartalsavgift och en hög avgift per markering .
70 INNEHÅLLER_KANSKESATS:     false
71 INNEHÅLLER_FUNDKANSKE_MEN_EJ_KANSKESATS:     false
72 ANVÄNDE_REINTRO:     false
73 FÖRSÖKER_ANVÄNDA_REINTRO:     false
74 XMLARR:     Den som ringer lite,ska,kunna,välja,att betala låg kvartalsavgift och en hög avgift per markering,.
75 PRIMÄRLEDMÖNSTER_ARR:     n,v,V,V,N,t
76 PRIMÄRLEDMÖNSTER_STR:     nvVVNt
77 PRIMÄRLEDMÖNSTER_OBJ:     [object Object]
78 array_av_xmlsatser:
79 array_av_spetsställda_enkelsatser:
80 array_av_frågor_för_meningen:
81 FÅR_IV:
82 ARRAY_AV_INGÅENDE_FLERORDSADVERBIAL:
83 INNEHÅLLER_TOMT_SEGMENT:     false
84 TILLHÖR_TESTMÄNGDEN:     false
85 has_HJALPFV:     true
86 ANTALFVEFTERLICENSIERING:     1
87 has_ETTENDAPRIMARFV:     true
88 ENDAPRIMARFVPLATS:     4
89 FVTYP:     ibland_hjv
90 is_FUNDAMENTLOS:     false
91 PRIMARVALENSVERB:     ska
92 is_HJALPVERBSKONSTRUKTION:     false
93 ANTALFINKONJ:     1
94 SLUTTECKEN:     .
```

Kodexempel 7 I Objektet `mening` initieras element och sätts allmänna värden typiskt på detta sätt. Uppställningen är en direkt tillståndsdump, utan förskönande. Alla möjliga attribut-värdepar initieras dock inte för alla textmeningar. Antalet attribut-värdepar som ingår har ökat konstant under arbetets gång och speglar de variabler som behöver sättas för att täcka olika konstruktionsfall. Som syns finns en viss redundans i innehållet. Denna listning är möjlig att visa i programmets gränssnitt. De flesta av attributnamnen antas vara självförklarande. (`RAPPFVARR`: vektor med finita anföringsverb.)

Det finns även ett stort antal globala variabler varav många sätts per analys av s-enhet och lika gärna skulle kunna placeras i ovanstående strukturer, vilka som nämnts, byggs ut kontinuerligt och är de strukturer som den stegvisa syntaxanalysen, enligt föregående kapitel, använder sig av.

4.1.2 Objektet *MR*

På grund av att den mer fullständiga analysen av meningen är en beskrivning av en struktur av obegränsad komplexitet och längd, används för den en dynamisk objektstruktur *MR* (’meningsrepresentation’) som mer utnyttjar den kapacitet ett objektbaserat programmeringsspråk tillhandahåller. Objektmodellen *MR* i programmet representerar hela meningen, huvudsatser, primära finita verbfraser, fält och chunkar i en struktur som programmatiskt kan undersökas, eller användas i sökuttryck, med hjälp av punktnotation. MR är toppobjektet i en struktur som kan beskrivas som trädformad, men det är viktigt att poängtera att det inte rör sig om en frasstrukturell analys som direkt speglar sådana konstituenter. I likhet med objektet *mening* har denna representation med ingående delar varit i ständig förändring under arbetets gång. I Nedanstående figurer är somliga understreck borttagna för läsbarhetens skull.



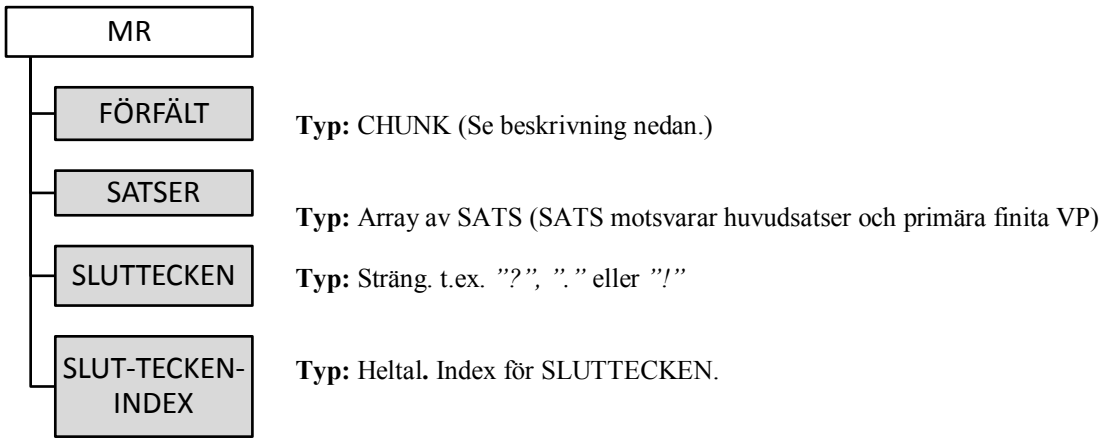
Objekthierarkin i det senare steget av analysen i programmet har ett toppobjekt *MR* (*meningsrepresentation)*, som underordnar strukturer av typen SATS, samt förfältsinnehåll (som ej hanteras som en del av de ’inre’ huvudsatserna.

Objektet *SATS* står för huvudsatser men även primära samordnade verbfraser. Dessa objektstrukturer kräver precis ett primärt finit verb. Ett samordnat primärt finit som *drack* i *Han åt och drack* utgör en minimal sådan SATS-struktur.

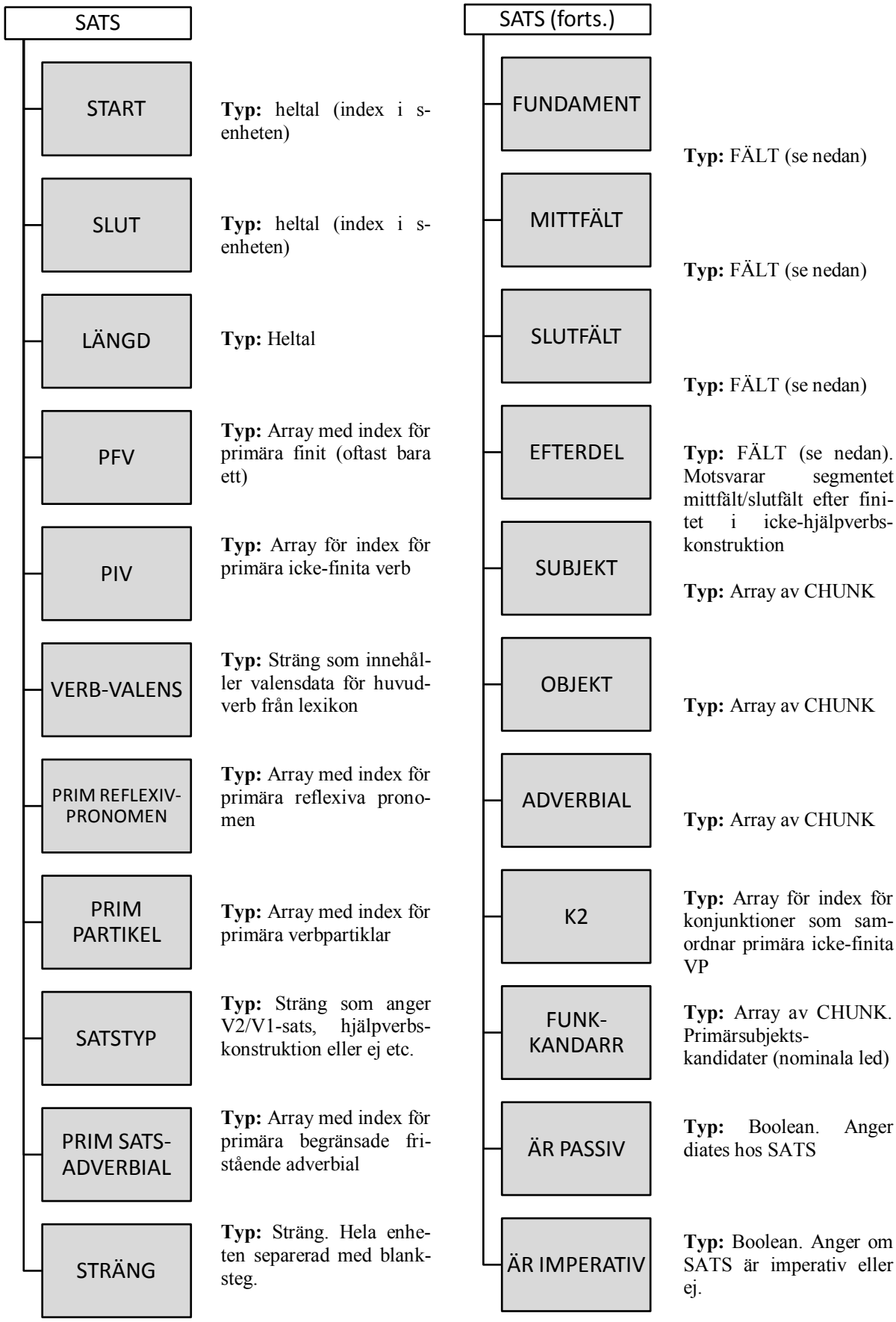
Objektet *FÄLT* representerar fältinnehåll i en SATS (huvudsats eller samordnad primär finit verbfras) enligt satsschemat förutom primära verb. Som sådana fält räknas här *fundament, mittfält* i hjälpverbskonstruktion, *slutfält* i hjälpverbskonstruktion eller *det sammanlagda ”fältet”* efter finitet i en icke-hjälpverbskonstruktion. (I den aktuella representationen som skiljer tydligt på hjälpverbskonstruktioner och andra, kallas detta sammanlagda fält för *efterdel*). De verb som avgränsar fälten i Diderichsens schema ingår som nämnts här inte i själva fälten.

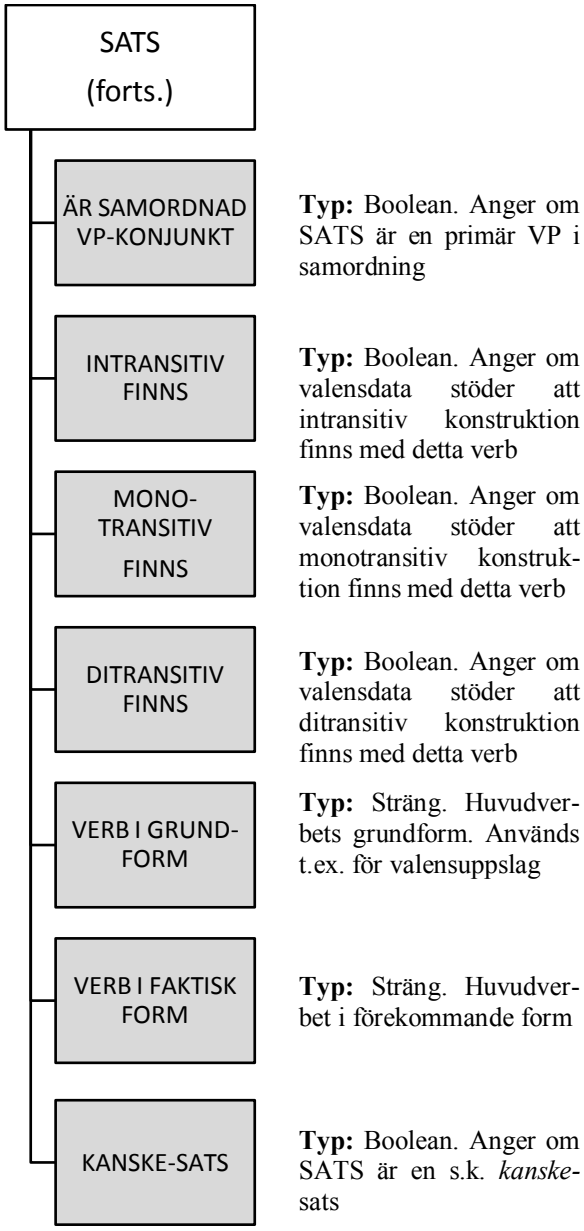
Objektet *CHUNK* beskriver uttömmande innehållen i *FÄLT*. Dessa strukturer är de segment som skapas genom den rangbaserade chunkningen men också de segment som lämnas orörda av denna – t.ex. interpunktioner, dessa blir inledningsvis egna strukturer av typen CHUNK. De olika segmenteringsstegen (resultatet av chunkningen samt de tre segmenteringsstegen) representeras genomgående som vektorer (endimensionella fält) av CHUNK.

Figur 23 Den hierarkiska strukturen inordnar objekttyperna enligt ovan. Egenskaper hos *MR, SATS, FÄLT* och *CHUNK* beskrivs i nedanstående figurer. Objektet MR med underordnade objekt ger en trädstrukturerad beskrivning av en mening. Det är viktigt att poängtera att detta inte är ett frasstruktursträd eller liknande.

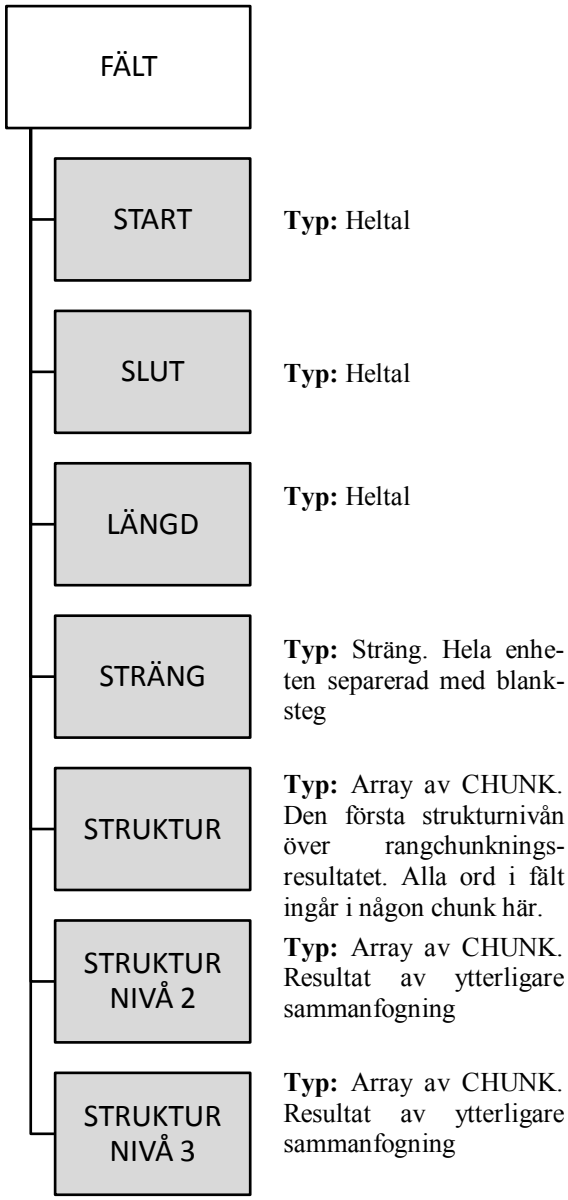


Figur 24 Det överordnade objektet MR är framförallt en behållare åt analysen som representeras huvudsatsvis.

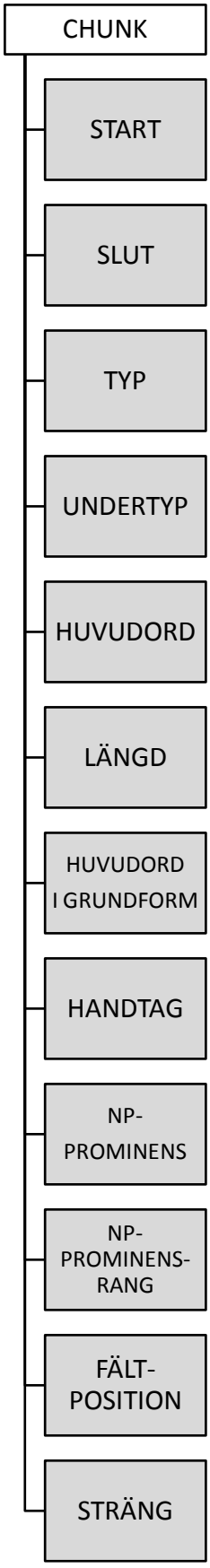

SATS
START
Typ: heltal (index i s-enheten)
SLUT
Typ: heltal (index i s-enheten)
LÄNGD
Typ: Heltal
PFV
Typ: Array med index för primära finit (oftast bara ett)
PIV
Typ: Array för index för primära icke-finita verb
VERB-VALENS
Typ: Sträng som innehåller valensdata för huvudverb från lexikon
PRIM REFLEXIV-PRONOMEN
Typ: Array med index för primära reflexiva pronomen
PRIM PARTIKEL
Typ: Array med index för primära verbpartiklar
SATSTYP
Typ: Sträng som anger V2/V1-sats, hjälpverbskonstruktion eller ej etc.
PRIM SATS-ADVERBIAL
Typ: Array med index för primära begränsade fristående adverbial
STRÄNG
Typ: Sträng. Hela enheten separerad med blanksteg.
SATS (forts.)
FUNDAMENT
Typ: FÄLT (se nedan)
MITTFÄLT
Typ: FÄLT (se nedan)
SLUTFÄLT
Typ: FÄLT (se nedan)
EFTERDEL
Typ: FÄLT (se nedan). Motsvarar segmentet mittfält/slutfält efter finitet i icke-hjälpverbskonstruktion
SUBJEKT
Typ: Array av CHUNK
OBJEKT
Typ: Array av CHUNK
ADVERBIAL
Typ: Array av CHUNK
K2
Typ: Array för index för konjunktioner som samordnar primära icke-finita VP
FUNK-KANDARR
Typ: Array av CHUNK. Primärsubjektskandidater (nominala led)
ÄR PASSIV
Typ: Boolean. Anger diates hos SATS
ÄR IMPERATIV
Typ: Boolean. Anger om SATS är imperativ eller ej.



Figur 25 Objekttypen *SATS* representerar huvudsatser, men också primära finita verbfraser. Skillnaden mellan dessa är att de senare saknar subjekt och fundament.



Figur 26 Objekttypen *FÄLT* representerar områdena som avgränsas av de primära verben, som alltså inte ingår i fälten här, och satsens gränser. De olika fälten är fundament, mittfält, slutfält och efterdel.



Typ: heltal (index i s-enheten)

Typ: heltal (index i s-enheten)

Typ: Sträng som anger NP, PP, bisatsinledare, licensierat/primärt finit/infinit verb, (lic/prim) adverbial, (lic/prim) partikel, (lic/prim) reflexivpronomen, övrigt icke-finit verb, konjunktion

Typ: Sträng som anger undertyp *hos de flesta chunktyper förutom PP*. Adverbfras, participfras, adjektivfras, icke-finit verb, pronomen, infinitivmärke, egennamn, tidsnomen (t.ex. *kväll* – möjliggör adverbialfunktion).

Typ: Sträng. Huvudord, vilket ofta innebär det sista ordet om det är ett segment som kommer direkt från rangchunkningen, dvs. saknar efterställda attribut. Obs: I PP räknas rektionens huvudord som huvudord. Dvs. *procent* i *Runt två procent*.

Typ: Heltal

Typ: Sträng (används för valensuppslag m.m.)

Typ: Sträng. Prepositionen i en PP (dvs. egentligen dess huvudord).
Dvs. *Runt* i *Runt två procent*.

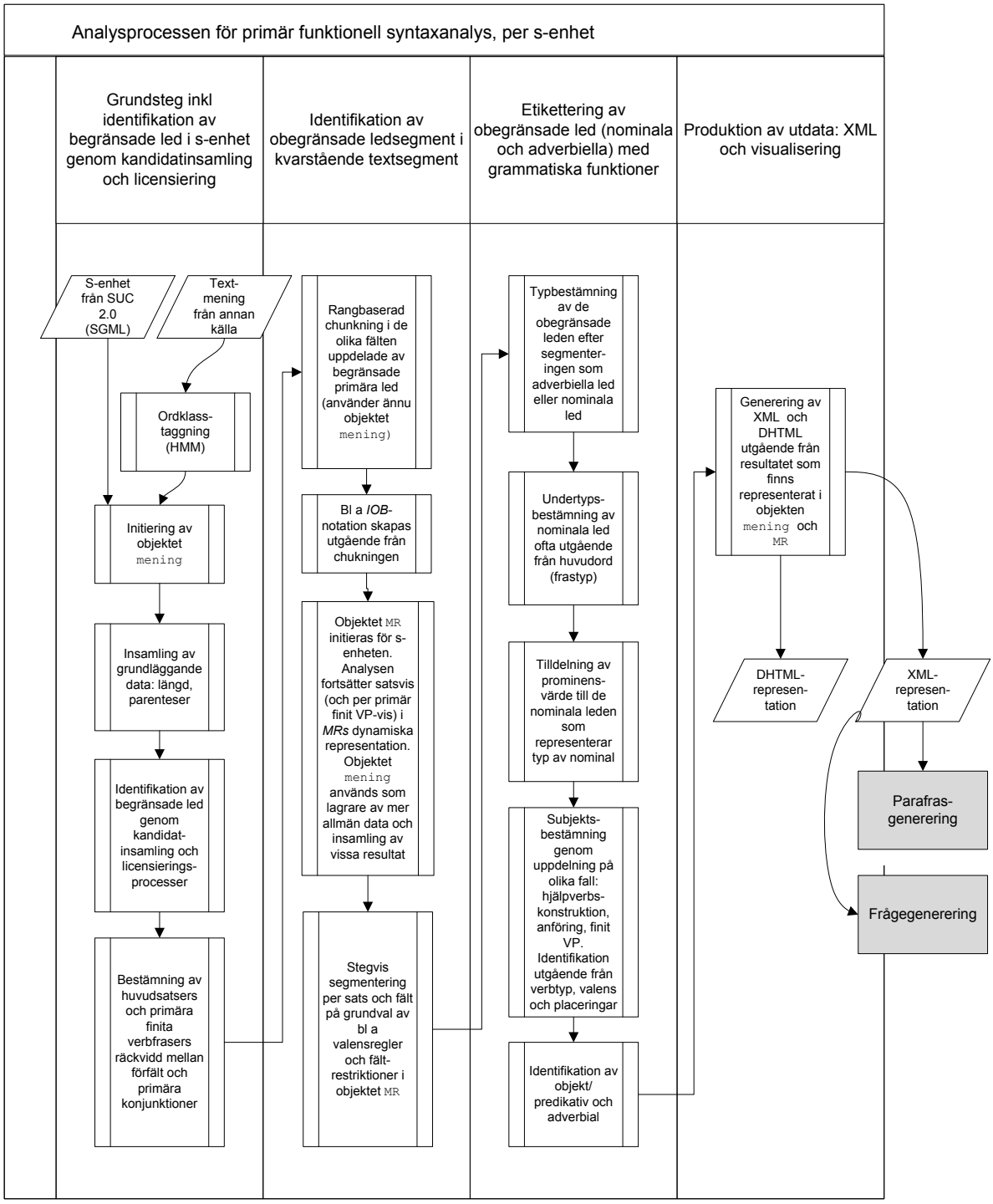
Typ: Sträng. En typ i allmänhet grundad på huvudordet i en NP, speciella kategorier för kasusmarkerade (subjekt/objekt) pronomen, personnamn, tidsnomen (t.ex. *kväll* – möjliggör adverbialfunktion), adjektiv i superlativ.

Typ: Heltal. En rangsiffra som ges åt ett segment på grundval av NP-prominens för bestämning av subjekt i vissa konstruktionstyper.

Typ: Sträng. Angivelse av Fundament, Mittfält, Slutfält eller Efterdel (dvs. positionen efter finitet i icke-hjälpverbskonstruktion).

Typ: Sträng. Hela strängen separerad med blanksteg: *En nyare , snabbare bil köptes* .

Figur 27 Objekttypen *CHUNK* används för att representera segmenten som är resultatet från den rangbaserade chunkning, segmenten i den stegvisa sammanfogningen samt de obegränsade funktionella leden i resultatet, subjekt, objekt/predikativ och adverbial.



Figur 28 Från indata till utformat används i inledningen främst objektet `mening` varefter objektet `MR` senare innehåller mycket av den optionella och dynamiska satsrepresentationen där t.ex. subjektsidentifikationen sker. De två tillämpningar som beskrivs i kapitel 5 använder XML-formaterad analys som indata.

4.1.3 Sökning mot *SUC 2.0* med objektrepresentationerna

I programmet har byggts in en sökfunktionalitet som möjliggör att finna exempel på olika konstruktionsslag, som för avhandlingens exempel, och fel i analysen. Detta är en funktionalitet som inte bygger på en förindexering utan slumpar fram s-enheter och samlar de som matchar uttryckta kriterier.[66] Det är denna funktionalitet som använts för att ge frekvensuppskattningar och vissa exempel i föregående kapitel. Genom de sökmönster som är möjliga, nämligen kombinationer från alla attribut-värdepar och på andra sätt härledda fakta om en textmening från den inre representationen, möjliggörs mycket komplexa sökmönster mot SUC 2.0. Sökmönstren kan därmed röra grundaspekter givna direkt i SUCs data och/eller aspekter från parsningen – dvs. satser, funktionella led, fält, fraser osv. Sökmönstret är i ren Javascriptkod mot objektmodellerna och kräver i nuvarande utförande denna kännedom om hur analysen representeras internt.

Sökning sker mot de inre representationerna av en textmening (*mening* och *MR*). Sökvillkoren kan kombineras med varandra och uttrycka i princip all information som används under analysen och i resultatformen. Sökvillkoren formuleras i Javascriptkod enligt Tabell 36. Sökning efter mönster har också kunnat användas för att finna felaktig taggning i SUC. Felaktig taggning kan nämligen ofta ge upphov till märkliga satsanalyser. Ett exempel är hur bisats- och relativsatsinledare som taggats som något annat kan upptäckas eftersom de kan resultera i en satskonjunkt med två icke-licensierade finit.

[66] Programmet som helhet bygger som nämnts ursprungligen på ett tidigare sökgränssnitt mot korpusen, som använde en indexering och därmed gick snabbare. Detta gränssnitt var visserligen också inriktat mot att finna s-enheter men sökkriterierna var direkt information från SUC och t.ex. fundamentlängd (dvs. bara viss enklare syntaktiskt härledd information). I programmet finns också möjlighet att söka efter s-enheter i SUC som innehåller ett specificerat graford och möjliga ordklasstaggningar för ett graford. Ett annat program som hör till projektet kan visa alla möjliga ord med en viss ordklasstagg (inklusive alla bestämningar). För fri sökning i denna korpus m.fl. rekommenderas *Språkbankens* webbgränssnitt. Aktuell adress: http://spraakbanken.gu.se/

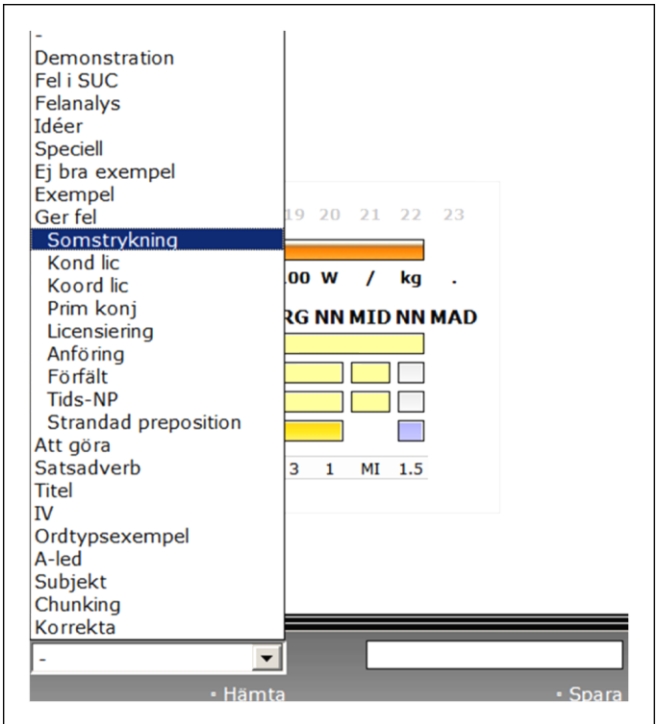
Meningslängd	`meningen.LEN == 7`
Det finns minst en huvudsats (eller primär finit VP)	`mr.satser.length >= 1`
NP i fundamentet	`mr.satser[0].FUNDAMENT.STRUKTUR_NIVÅ_3[0].TYP == "NP"`
Fundamentlängd (i första satsen)	`(mr.satser.length>=1) && (mr.satser[0].FUNDAMENT.LÄNGD == 4)`
Sluttecken	`mr.SLUTTECKEN == "!"`
Förekomst av ditransitivt verb i första satsen	`(mr.satser.length>0) && (mr.satser[0].DITRANSITIV_FINNS == true)`
Textmeningar med minst två primära satskonjunkter eller primära finita verbfraser (vilka båda kallas sats i koden) där de två första har samma fundamentlängd av minst 2 ord och där strängen *på* finns någonstans i textmeningen. `Stringify` ger en strängrepresentation mellan två positioner av enheten, och ger här hela meningen.	`(mr.satser.length > 1) && (mr.satser[0].FUNDAMENT.LÄNGD > 1) && (mr.satser[0].FUNDAMENT.LÄNGD==mr.satser[1].FUNDAMENT.LÄNGD) && (is_in_string("på", stringify(meningen, 0, meningen.LEN - 1).toLowerCase())`

Om en enkel grammatikbeskrivning skulle förutsätta att nominalfras i ett mittfält utgör satsens subjekt kan ovanstående kod användas för att bekräfta om det finns en sådan. (I Javascript innebär, som i C, ”=” tilldelning av värde medan ”==” är det boolska värdet för om ekvivalens gäller.)	`var mittf = mr.satser[0].MITTFÄLT;`
	`var det_finns_nom_struktur_i_mittfältet = false;`
	
	`for (i=0; i<mittf.STRUKTUR.length; i++)`
	`{`
	`if (mittf.STRUKTUR[i].TYP == ”NP”)`
	`{`
	`det_finns_nom_struktur_i_mittfältet = true;`
	`}`
	`}`

Tabell 36 Några exempel på sökuttryck och andra uttryck som kan användas i programmet. Som nämnts går det att använda dessa för mycket specifika villkor. Det som då skiljer mekanismen från andra söksystem mot bl.a. SUC är möjlighet till sökning med uttryck i syntaxanalysen. Ett kodexempel finns även i Appendix. (*Stringify* ger ordsträngen mellan två ordindex.)

Övrig funktionalitet i implementationen

Programmet innehåller ett system för lagring och hämtning av påträffade s-enheter (huvudsakligen från SUC) i för närvarande 25 kategorier. Dessa kategorier inbegriper olika feltyper, bl.a. för att spåra fel i licensiering, och exempelkategorier för att spara skilda typer av påträffade enheter, vilket innebär att den kontinuerliga analysförbättringen kunnat ske med avseende på en samling exempel för olika konstruktionstyper. Vidare finns en kategori för insamlandet av vad som antas vara fel i SUC 2.0, insamling av vissa flerordskonstituenter, exempel för att förbättra sammanslagning/åtskillnad av chunkar och identifikation av funktionella roller, enligt Figur 29.



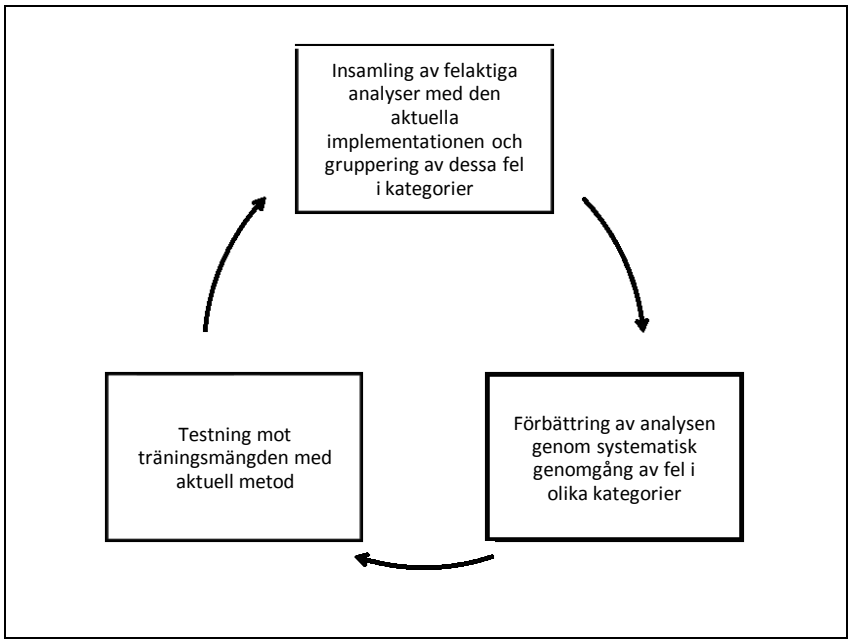
Figur 29 S-enheter sparas i kategorier tillsammans med kommentarer i gränssnittet. Dessa grupper som används för att systematiskt göra exempelgrundade förbättringar inbegriper för närvarande feltyper för olika analysdelar som licensiering och antagna fel i SUC.

4.2 Det praktiska arbetet med analysförbättring

Implementationen som inbegriper syntaxanalys är inte enbart en spegling av en färdigbyggd analysator utan är i hög grad gjord för förbättra analysen dynamiskt vid åsyn av nya textmeningar. Samtidig förbättring och insamling av exempel och fel – även fel i själva SUC-märkningen – har skett under en lång period och ett stort antal s-enheter från träningsmängden har analyserats för att samla dessa

i kategorier för att metodförbättring genom regelomskrivning ska kunna ske på grundval av många exempel, enligt Figur 30.

Utvecklingen av programmets analys har i hög grad antagit formen av felrättning av gjorda analyser med den regeluppsättning som använts för tillfället. Nya regler har iterativt skapats för att korrekt täcka fler fall, mycket ofta rent programmeringstekniskt på formen *undantag från huvudregeln* eller *undantag från undantag* etc.



Figur 30 Gemensamt för arbetet med förbättring av de olika delarna av implementationen är den manuella, iterativa process med insamling av s-enheter i SUC som exemplifierar olika felaktiga analyser. Som synes sker både testning och förbättring med underlag från *träningsmängden* (dvs. från ca 89 % av SUC 2.0). *Testmängden* har sparats till mer formella resultatberäkningar.

I själva gränssnittet finns även möjlighet att se och modifiera de viktiga ordmängder som insamlas under utvecklingen[67]. Det finns en högerklicksmeny för att snabbt lägga till ett ord i en analyserad mening i olika ordmängder, och detta medger omedelbar lexikonutökning och kontroll av den nya analysen. Den kontinuerliga förändringen av ordmängderna är alltså relativt enkel från ett ”manuellt träningsperspektiv”. En sorts allmän målsättning för projektet har varit att nå ett utvecklingstillstånd där syntaxreglerna är så gott som fastställda och enbart

[67] I skrivande stund gäller detta förnamn, efternamn, (finita) hjälpverb, mängdord (*liter, kg*), personattribut (*herr, finansminister*), adverb som modifierar prepositionsfraser när de står före (*upp, direkt*), adverbattribut (*hem, dit*) och huvudord i NP som möjliggör adverbialfunktion (*gång, tag, sommaren*).

delar av den närmast oändliga uppgiften ordmängdsutökning kvarstår, dvs. de syntaktiska reglerna som hanterar t.ex. förnamn i svensk text är klara medan gruppens samtliga ord inte är insamlade.

Allmänt om implementationstekniken

Programmet skulle kunna göras mycket snabbare omskrivet i programkod som kompileras istället för att interpreteras som är fallet här. Det är okänt vad metoden som sådan har för tidskomplexitet men troligen är den långsammare än de flesta.[68] Den kraftfulla inre representationen av textmeningar (se föregående avsnitt) är heller inte optimerad när det gäller tidsåtgång. Rent praktiskt är dock systemet tillräckligt snabbt för uppbyggandet av en analysprocedur. Den aktuella implementationen är gjord i *Jscript*-kod (Microsofts version av Javascript) vilket innebär ett snabbt och flexibelt sätt att arbeta. Tekniskt sett är koden inrymd i en s.k. *HTA* (*HTML Application*), vilket grovt motsvarar en webbläsare (Internet Explorer på Windows) men med vissa friheter. Dessa är att programmet kan köras lokalt utan att behörighetsprompter stör och att Jscript på Windows exponerar *FileSystemObject,* vilket dessutom ger tillträde till systemets filsystem med läs- och skrivrättigheter, vilka används flitigt. Applikationstypen kan beskrivas som en exekverbar fil rättighetsmässigt, men bygger på ickekompilerad scriptkod. Förutom *FileSystemObject* har vissa andra Windows-specifika ActiveX-kontroller använts (t.ex. datastrukturen *Dictionary*) men sedermera skrivits om för att möjliggöra en eventuell, framtida mer systemoberoende Internet-version. Detta är rimligen den enda fritextparsern för svenska i detta C-liknade skriptspråk. Programmeringsstilen är funktionsbaserad tillsammans med objektbaserad programmering för de nämnda strukturerna.

4.3 Ordklasstaggning i systemet

För att åstadkomma den beskrivna analysen av svensk text i implementationen har framförallt SUC 2.0 använts, men även fri text. Inledningsvis var programmets ordklasstaggare för fri text en jämförelsevis lågpresterande bigrambaserad dold Markov-modell (HMM) med Viterbi-algoritmen (1967) med en enklare suffixanalys för taggning av okända ord. Därför har inledningsvis framförallt s-enheter från SUC varit manuell träningsdata under utvecklingen. Senare förseddes implementationen istället med en motsvarande trigrambaserad ordklasstag-

[68] På en normal, inte helt ny, PC tar parsning utan ordklasstaggning och visualisering för närvarande ca 5–10 hundradelar per s-enhet under en sökning i gränssnittet. Helt programoberoende förändringar som hårdvaruförändringar och versionsuppdatering av ramverk för HTA har dock haft minst lika stor betydelse för snabbheten som viss genomförd kodoptimering.

gare med mer utarbetad suffixstatistik för taggning av ej lexikoninförda ord istället. Nedan redovisas mätdata för denna ordklasstaggare, programmerad i Javascript. [69]

4.3.1 Trigrambaserad ordklasstaggning i systemet

2007 färdigställdes en HMM-baserad trigramtaggare med Viterbi-algoritmen i Javascript. För tillämpningarna i kapitel 5 används denna ordklasstaggare. Källan som främst använts för arbetet med taggaren är den ingående beskrivningen i Carlberger och Kann (1999) med sannolikhetsvärden utgående från Formel 6 och Formel 7. Förutom att implementationssättet är annorlunda jämfört med i artikeln, finns även ett antal metodskillnader. Istället för *interpolär smoothning* används *additiv smoothning* som i Nivre o.a. (1996), dvs. i uppskattningen av sannolikheter antas ett antal förekomster för varje bi- och trigram – även de som ej förekommit i träningsdata – på vilket antalet faktiska förekomster adderas. Korrektheten har utvärderats till 95,3 % för alla ord, se Tabell 38.[70] Den är inte en del av metoden som sådan och skulle med fördel kunna bytas ut eller förbättras. Om jämförelser görs angående syntaxanalysens korrekthet och användbarhet för fri text i kommande kapitel är det en sorts poäng att goda resultat i alla fall inte beror på själva taggningen.

Formel 6 Kontextuell sannolikhet

Kontextuell sannolikhet: $P(c_i \mid c_{i-2}\ c_{i-1})$, dvs. sannolikheten för en tagg givet två föregående taggar, uppskattas som

$$\frac{(f(c_{i-2}, c_{i-1}, c_i) + k_1) / (n + c^3)}{(f(c_{i-1}, c_i) + k_2) / (n + c^2)}$$

där n = antalet löpord, c = antalet olika ordklasser, f = frekvens

- Division med (n + c²) samt (n + c³) har ignorerats i implementation för att spara tid.
- Bästa korrekthet nåddes med $k_1 = 0{,}2$ och $k_2 = 200$.

[69] Detta utgjorde ett ganska långdraget arbete. Speciellt tackas Leif Grönqvist, Viggo Kann och Harald Hammarström för hjälp med konstruktionen av de två ordklasstaggarna.

[70] En synpunkt är att tester i SUC, där en annan delmängd av SUC har använts för konstruktion faktiskt borde innebära att vokabuläret i kanske alltför hög grad är känt, p.g.a. att test- och träningsdatan kommer från samma texter. I både Carlberger och Kanns system och i aktuell implementation ligger andelen okända ord på omkring 4–7 %.

Formel 7 Lexikal sannolikhet

Lexikal sannolikhet uppskattas som kvoten av antalet ordförekomster med viss tagg dividerat med frekvensen för taggen.

P (ord | ordklasstagg) uppskattas som

$$\frac{\text{frekvens (ord med ordklass)}}{\text{frekvens (ordklass)}}$$

där frekvensen för ordet med aktuell ordklass är minst 1.

En del i implementation som är originell och som också leder till snabbare analys men som framförallt har som syfte att öka korrektheten var att diskriminera vissa ord-taggpar enligt nedanstående tabell. Det kan vara t.ex. egennamnstagg för ord med liten bokstav etc. Följande ”POS-diskrimineringsregler” användes huvudsakligen i de resultat som presenteras i
Tabell 37.

Kategori	Exempel
Allmänt förbjudna ord-tagg-par (ofta p.g.a. fel i SUC)	*”sticks” med taggen VB-PRT-SFO*
Ej utländska ord (ord som någon gång i SUC markerats som utländska men som i princip aldrig bör vara det i svensk text)	*Och, med, som*
Ej presensverb om föregående ord tillhör en mängd typiska presens hjälpverb	*”skulle får”*
Ej infinitiv om föregående ord är ”som”	*”som få”*
Ej egennamn om ordet ej har stor begynnelsebokstav	*”per”*
Ej något annat än egennamn om egennamn är en möjlig tagg och ordet ej finns först i mening.	*…var Sally…*
Ej någon annan tagg än ”PN” för ordet ”mycket” om ”av” följer	*”…mycket av allt”*
”Att” ska ej vara infinitivmärke (IE) om finit verb följer	*”…att springer”*

Tabell 37 De diskrimineringsregler som använts är ett medel för att undvika olika ordtaggkombinationer redan i själva HMM-algoritmen. Genom att omöjliggöra vissa kombinationer beroende på kontext (som vid detta laget ännu inte är taggad) kunde vissa fel undvikas.

Resultatmässigt har taggaren slutgiltigt utvärderats till korrekthetsvärdena i Tabell 38. Korrektheten är 95,3 % för alla ord. Det finns några grundläggande skillnader mellan förutsättningarna här och dem som fanns i Carlberger och Kann (1999). I den aktuella implementationen har felanalysen för att förbättra

korrektheten uttryckligen grundats på (delmängder av) testmängden, detta kan innebära att resultatet låter påskina att taggaren är bättre än vad den egentligen är. Det nedanstående är resultat från test på precis hela testmängden. En annan gynnsam faktor här, jämfört med i Carlberger och Kann, är att s-enheter i SUC vari misstänkta feltaggningar påträffats tagits bort helt (de har i testningen tagits bort från testmängden, men de fanns i träningen).

	Carlberger och Kann (1999)	Aktuell implementation
Antal utvärderade meningar (s-enheter) i aktuell testning	1006	3692
Antal ord	16 378	57 266
Korrekt taggade meningar (s-enheter)	58, 6 %	Ej utrett
Korrekthet för ord i lexikonet	96,6 %	96,0 %
Korrekthet för ord som saknas i lexikonet	92,0 %	82,6 %
Korrekthet för alla ord	**96,3 – 96, 4 %**[71]	**95,3 %**

Tabell 38 Resultaten visas här i jämförelse med C-implementationen av Carlberger och Kann (1999) för att visa att resultaten inte är alltför långt därifrån, även om förutsättninga har varit olika.

Jämförelsen i Tabell 38 är mest ett sätt att visa att korrektheten är i närheten av den i den pedagogiska artikel som utgjorde förebilden. Som redan nämnts är det rimligt att ett framtida system för hantering av fri text byter ut aktuell ordklass-taggning.

[71] Korrektheten blir där 96,4 % om även *SAOL* (1986) används för att utöka lexikonet.

4.3.2 Om betydelsen av fel i ordklasstaggningen

En viktig aspekt angående jämförelsen mellan ordklasstaggarna är att de är tänkta att fungera i olika system. För det aktuella systemets uppgift är det ofta stor skillnad i hur menliga olika taggningsfel är. Här har många feltyper som handlar om felaktiga 'undersärdragsvärden' hos korrekt taggade *ordklasser* mindre betydelse på grund av rangchunkningens och uteslutningsmetodens relativa okänslighet, en skillnad jämfört med ett granskningsprogram. Faktum är att fel i ingående särdragsvärden hos korrekt taggade ordklasser i NP/PP utan efterställda attribut (adjektiv, particip, adverb, substantiv m.fl.) ofta är betydelselösa eftersom dessa ordklasser i syntaxanalysen senare bara representeras med en rang som står för ordklassen. Däremot är feltaggning av/till bisatsinledare och av/till verb potentiellt förödande eftersom analysen så tydligt beror på att inledande primärfinitsidentifikation blir rätt. P.g.a. detta har en speciell kvot för antal fel där bisatsinledare eller verb är inblandade (dvs. de har missats eller satts in fel) per testade s-enheter mätts i testerna. Denna kvot ligger i testerna på 0,18. Det ska nämnas att betydelsen av denna kvot inte ska övertolkas eftersom t.ex. sammanblandning av två olika bisatsinledande taggar inte leder till fel medan somliga till synes ofarliga fel faktiskt kan ge fel i syntaxanalysen. En mindre analys av en delmängd av felen i en undersökning med avseende på allvarighetsgrad för den aktuella tillämpningen ('schemabaserad syntaxanalys') visar hur ungefär hälften av alla fel förhoppningsvis inte behöver påverka uppgiften, enligt Tabell 40 och Tabell 39.

Det måste klargöras att det ovanstående försöket till skattning av allvarligheten i feltaggningen hos olika taggfel är grov och alltså kan vara missvisande i enskilda fall. Till exempel kan felaktiga mindre centrala särdragsvärden hos substantiv ge upphov till felaktig *som*-strykning, medan feltaggning av verb kan vara ofarlig, t.ex. när dessa ord finns på underordnad licensierad satsnivå. Den verkliga effekten ordklasstaggningen och tillhörande fel kan ha, syns i resultatet för syntaxanalysen där det kan undersökas i vilken grad slutliga fel härrör från olika delsteg av hela parsningen.

”Ofarliga fel” (0)	55 %
Allvarliga fel (F)	16 %
Andra fel (1)	30 %

Tabell 39 En sammanfattning av feltyper i en mindre felmängd (208 fel på formen enligt Tabell 40) med avrundade värden visar att många fel kan vara harmlösa.

Fel	Ord	Utsatt tagg	Facit	S-enhet
0	**Vad**	*HP-NEU-SIN-IND*	*HA*	aa01a-059 Vad Israel gäller har både den saudiska varningen och utvecklingen i Irak vederbörligen uppmärksammats.
F	**Ökat**	PC-PRF-NEU-SIN-IND-NOM	*VB-SUP-AKT*	aa02c-013 Men om åldersstrukturen hade varit enda orsaken till sjukvårdskostnadernas ökning skulle dessa ha ökat med 13 procent mellan 1970 och 1985 .
0	**Stegs**	NN-NEU-SIN-IND-GEN	NN-NEU-PLU-IND-GEN	aa02d-005 Soldaterna stod med 30 stegs lucka och körde iväg alla fotgängare .
0	**Ett**	DT-NEU-SIN-IND	RG-NEU-SIN-IND-NOM	aa02d-025 Den 20 maj är det ett år sedan Li Peng med flera införde undantagstillstånd för att kväsa det växande upproret.
0	**Ut**	PL	AB	aa02d-045 Den diskussion som ändå ägt rum enligt vad som sipprat ut , har gällt förhållandet mellan centralmakten och provinserna .
0	**Ni**	PN-UTR-PLU-DEF-SUB	PN-UTR-SIN-DEF-SUB	aa03a-038 Jag vill utnyttja tillfället till att fråga om hur ni ser på relationen till Boris Jeltsin . "
F	**Levereras**	*VB-PRS-SFO*	*VB-INF-SFO*	aa03b-044 Den modernare Jaktviggen började levereras 1980 och detta sista planet som Lars Rådeström nu står i begrepp att lyfta med är det 149:e .
1	**Som**	KN	*HP------*	
0	**Ofarliga**	JJ-POS-UTR/NEU-PLU-IND/DEF-NOM	JJ-POS-UTR/NEU-SIN-DEF-NOM	aa04b-002 Tobaksindustrins gamla dröm från 60-talet om den ofarliga cigarretten håller på att bli verklighet .
0	**Land**	NN-NEU-PLU-IND-NOM	NN-NEU-SIN-IND-NOM	aa05a-025 Parallellt med detta lovade man ju hårda tag för att ge de jordlösa svarta land från storfarmarnas egendomar som täcker landets bästa jord; 90 procent av dessa storfarmer är vita .
0	**Års**	NN-NEU-PLU-IND-GEN	NN-NEU-SIN-IND-GEN	aa05b-031 Riksbanken fråntogs dessutom alla sina silvermynt , som sammanlagt uppgick till ett värde av 35 miljoner litas (värdet angivet i 1940 års priser).
0	**Osäkra**	JJ-POS-UTR/NEU-SIN-DEF-NOM	JJ-POS-UTR/NEU-PLU-IND/DEF-NOM	aa05b-051 - Talen är dock ytterst osäkra , eftersom ingen egentligen kan spekulera om vad som hade hänt om vi inte hade blivit en del av Sovjetunionen , säger Aleskaitis.
0	**Vad**	*HP-NEU-SIN-IND*	*HA*	
F	**Om**	PP	*SN*	aa05c-011 – Om någon delstat skall stå modell för en sovjetisk övergång från centralstyre till marknadsekonomi är det vår.
1	**Vår**	NN-UTR-SIN-IND-NOM	PS-UTR-SIN-DEF	
0	**Sovjetledare**	PM-NOM	NN-UTR-SIN-IND-NOM	aa05c-031 Den stod i skarp kontrast till det förra besöket av en Sovjetledare i delstaten .

Tabell 40 Den ovanstående tabellen är hämtad ur denna felanalys där mindre allvarliga fel markerats med 0, medelsvåra fel (med rimlig chans att leda till fel) markerats med 1 och grava fel (som angår bisatsinledare och verb) markerats med F. Understrykning i ordkolumnen betyder okänt ord. Kursivering i de andra ordklasskolumnerna markerar avgörande fel.

4.3.3 Fel och inkonsekvens i *SUC 2.0*

De utvärderingar som här görs mot SUC 2.0 har som syfte att bestämma korrekthet eller frekvens i *korrekt ordklasstaggad text*. Feltaggning i Stockholm Umeå Corpus är relativt sällsynt med tanke på storleken. Inte desto mindre innehåller den gott om fel att finna, vilket kan bli tydligt vid en längre tids syntaxanalys. I Johansson Kokkinakis (2003) avhandling om ordklasstaggning exemplifieras felen med följande exempel.

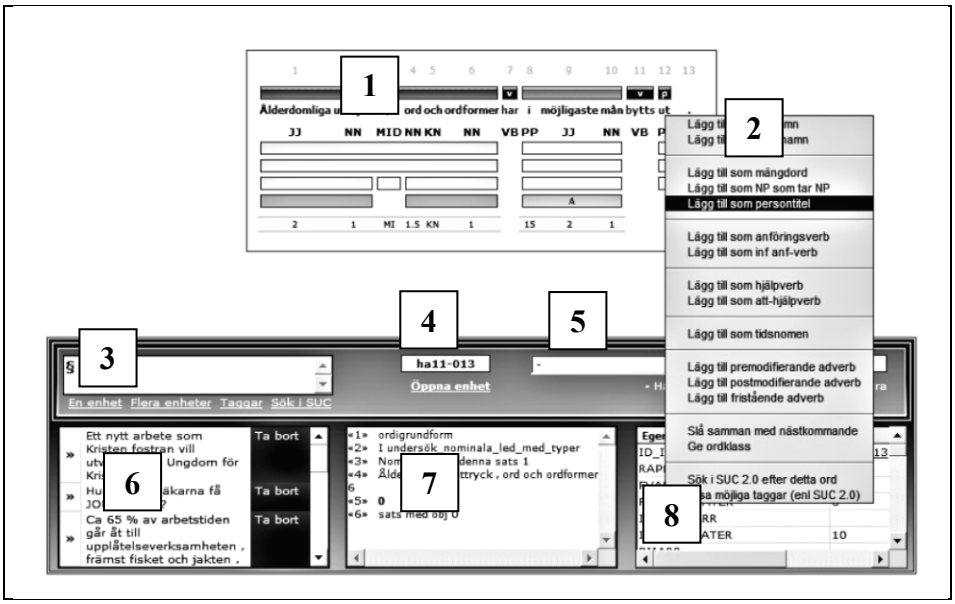
Ex 60 Det ger en körsträcka på upp till 70 kilometer och en toppfart på 110 km/h.
(ec05b-029)

Felet i Ex 60 är att *kilometer* taggats *substantiv-utrum-singular-obestämd form-nominativ*, när numerus borde ha varit *plural*. Dessutom exemplifieras inkonsekvens i taggningen genom att visa hur *vara* i uttrycket *ta till vara* på taggats olika i olika fall (ibland som verb, ibland som adverb). Detta är ett exempel av många där frågan om ordklasstilldelningen ställts på sin spets (det som här poängteras är dock *inkonsekvensen* i taggtilldelningen).

Carlberger och Kann (1999) antog att antalet felaktiga taggningar i deras version av korpusen var ca 5 000, dvs. 0,5 % av löporden. Därutöver identifierades inkonsekvent taggning, vilken i denna källa rättades till, och därmed åstadkom bättre underlag för en statistisk ordklasstaggare. I arbetet här med manuell testning mot den större delen av SUC (träningsmängden) märks speciellt de taggningsfel som ger tidiga grova syntaxfel i analysen – nämligen de som är felaktig taggning av/till finita verb och av/till bisatsinledare. Däremot leder faktiskt inte feltaggningen i Ex 60 ovan till fel i den aktuella parsningen. De s-enheter som antas innehålla fel kan plockas undan genom gränssnittet i programmet. I utvärderingen av korrekthet för olika aspekter undantogs alltså påträffade s-enheter med fel i SUC 2.0 när det var möjligt (både från tränings- och testmängd). Det samlade antalet undantagna s-enheter p.g.a. antaget fel under utvecklingen av den schemabaserade parsningen är, i skrivande stund, drygt 1150 s-enheter. Dessa kommer huvudsakligen från träningsmängden eftersom denna mängd analyserats och undersökts betydligt mer.

4.4 Beskrivning av gränssnittet

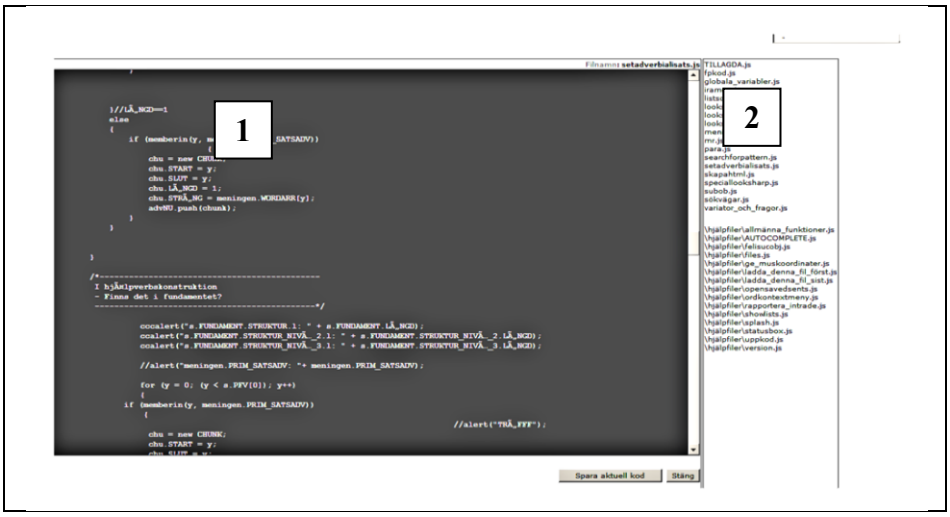
Programmets gränssnitt har förutom ren visualisering av analyserade enheter också en direkt roll i programförbättringen genom dynamisk uppdatering av ordmängderna. Detta sker genom speciella listningar med möjlighet att därifrån ändra programmets ordlistor över hjälpverb, persontitlar, anföringsverb, personnamn, huvudord för NP-formade adverbial etc. Dessutom finns möjlighet att direkt i en analyserad mening, via en kontextmeny (högerklicksmeny), lägga till ett påträffat ord och att undersöka vad förändringen leder till för konsekvenser för syntaxanalysen. Dessa gränssnittsfunktioner har möjliggjort omedelbar korrigering och kontroll av resultatet, vilket har varit viktigt för ett effektivt manuellt arbete där textmeningar, främst framslumpade sådana från SUC, använts för att testa och kontrollera systemet fortlöpande. De följande figurerna visar den viktigaste funktionaliteten.



Figur 31 I huvudgränssnittet används en färgkodning, som ej blir synlig i tryck, för att markera grammatiska funktioner ovanför texten och den stegvisa segmenteringen utgående från rangchunkningen (nederst i analysen).

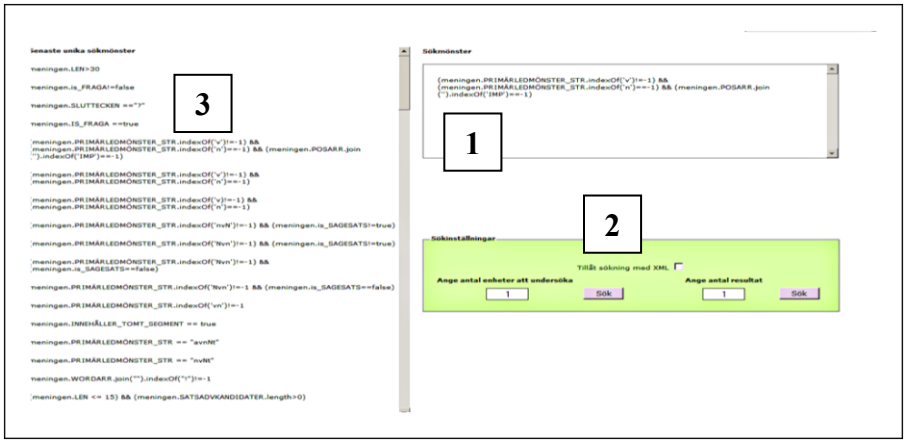
1) En syntaxanalys visualiseras med extrainformation som t.ex. övriga taggningssärdrag vid muspekning.
2) Kontextmenyn kan placera ett förekommande ord i ordmängder som förnamn, nomen med adverbialfunktion, söka övriga enheter från SUC 2.0 som innehåller ordet och ge SUCs alla förekommande taggar för ordet.
3) Ett textformulär möjliggör textinput (en eller flera textmeningar) för analys.
4) ID för aktuell s-enhet i SUC visas eller kan skrivas in för hämtning.
5) Kategorisystemet innebär insamling av enheter i olika kategorier, t.ex. olika felaktiga analyser och hämtning av dessa.
6) Uppvisning för återgång till inladdade s-enheter.
7) Statusmeddelanden för diverse information om analysen.
8) Utvalda attribut-värdepar från objektet *Mening* för aktuell enhet.

Figur 32 Nederst i huvudgränssnittet (se Figur 31) finns möjlighet att ändra analysens utseende, att evaluera kod mot den för närvarande inlästa s-enheten, att lägga till mindre koddelar explicit, att hämta ID för senast analyserade enheter samt att ladda minnet med SUCs s-enheter filvis, vilket ungefär fördubblar sökhastigheten (se Figur 34).

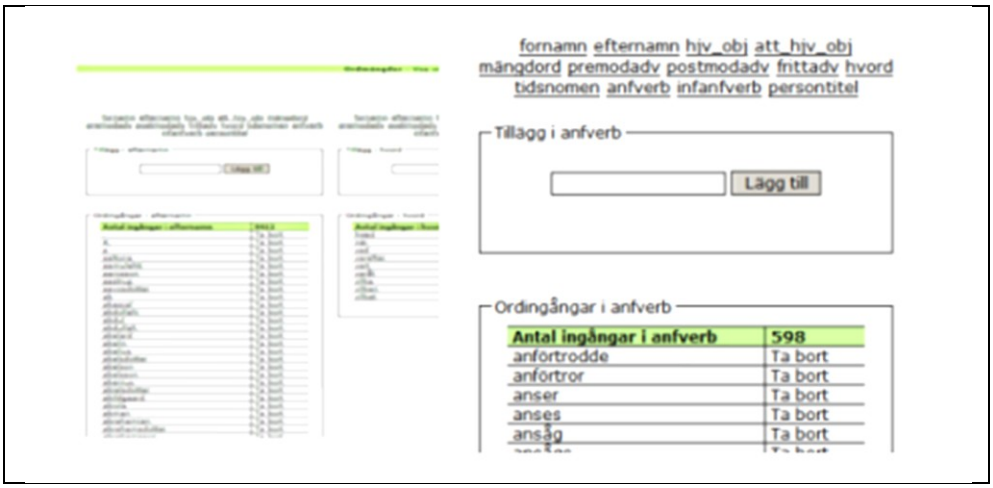


Figur 33 Under arbetets gång har flera olika editorer använts. I slutet byggdes en egen editor i själva programmet för att snabbt kunna ändra koden i de källfiler som programmet innehåller.

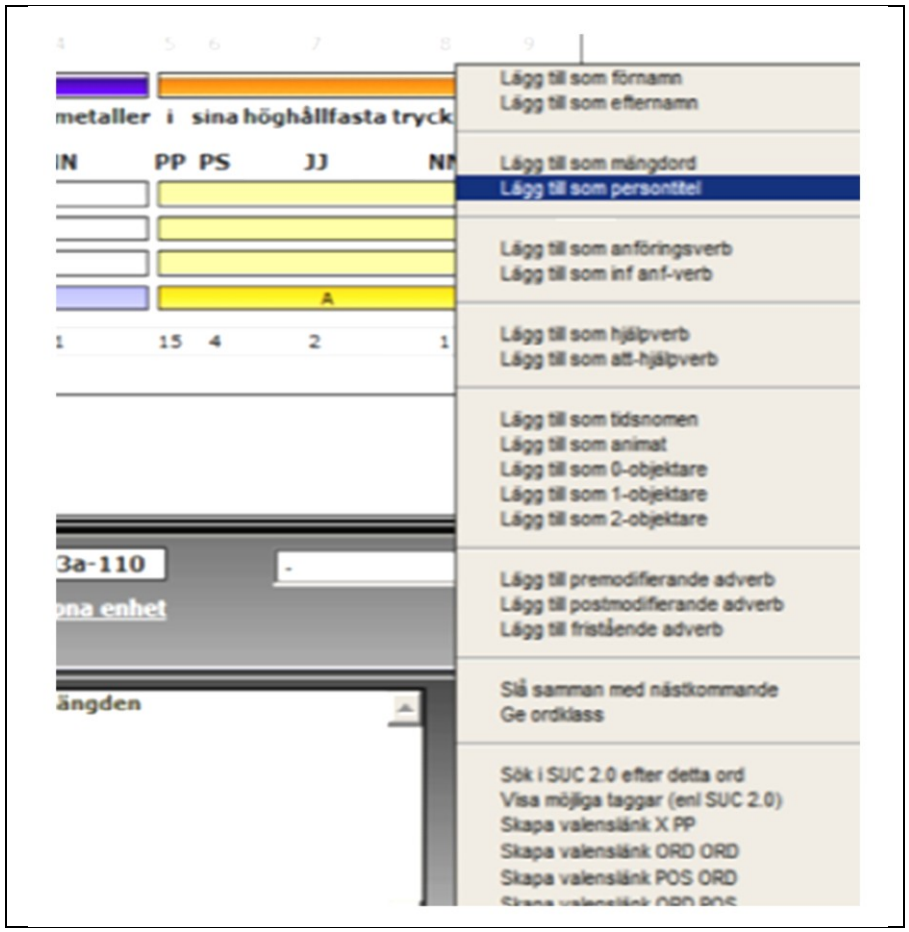
1) Editordel med möjlighet att ändra och spara programmets kodbas direkt.
2) Listning av skriptfiler som kan öppnas och modifieras.



Figur 34 Sökvyn är den vy där de ovan beskrivna sökuttrycken formuleras.
1) Sökuttrycksformulär
2) Sökspecificering: sökning i ett visst antal enheter (för statistik), eller sökning efter ett visst antal matchande enheter (för exempel)
3) Senaste unika sökuttryck, för snabb återanvändning



Figur 35 Ordmängdsvyn möjliggör uppvisning och modifiering av de ingående ordgrupperingarna, t.ex. persontitlar och mängdord (se Appendix), vilka också kan läggas till via kontextmenyn. Dessa kanske verkar ha en framträdande del av detta arbete men då bör det påpekas att dessa ordgrupperingar, med speciell syntaxfunktion som måste samlas för att täcka funktion utöver den information som finns i taggningen, på något sätt behöver finnas i parsningssystem för svenska.



Figur 36 Genom kontextmenyn möjliggörs direkt inplacering av ordet (den form det förekommer i) som förnamn, efternamn, mängdord, persontitel, hjälpverb, nomen med potentiell adverbialfunktion, animathetsklassificering, visst antal förväntade objekt (för precis den aktuella ordformen) samt fristående eller modifierande adverb. Från menyn går också att finna övriga enheter i korpusen som innehåller det aktuella ordet (funktionalitet från en tidigare indexering), visa vilka olika ordklasstaggar det förekommer med, samt skapa länkningsregler med det påföljande ordet på grundval av ord och/eller ordklasstagg hos de två involverade orden.

4.5 Viss användning av valensinformation

Följande avsnitt gäller framtagande av valensinformation för användning i analysen. Frågan har varit om det är rimligt att använda ordboksinformation för parsningen, och i så fall hur mycket arbete som krävs för att göra den användbar i detta sammanhang, samt vilken betydelse den får i praktiken för syntaxanalysen, dvs. hur ofta den är avgörande.

Valensinformationen som här omarbetats formatmässigt är en dyrbar resurs och en del av projektet som tydligt berikar programmet med manuellt utarbetad syntaktisk information. Det sågs tidigt som en central fråga hur väl denna sorts resurs kunde användas och förbättra analysresultat. Användningen av de aktuella källorna måste betraktas som osofistikerad i jämförelse med den informationsutvinning som kan göras av en läsare av de aktuella ordböckerna. Anledningen är huvudsakligen att grafiska ord ofta svarar mot flera lemman och lexem och betydelsedisambiguering skulle krävas mycket ofta. Omarbetningen från ordboksformen var tidsödande när det gällde att undersöka om det gick att omformulera dessa exakt till programmerade regler för sammanlänkning av segment.

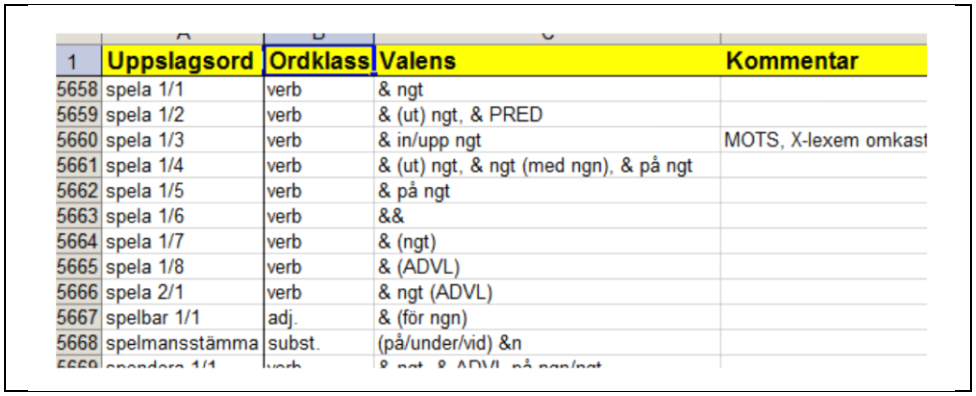
Trots att den information som kan erhållas från ett medium som är avsett för mänsklig läsning förlorar finare betydelsedistinktioner, så innebär lexikonets storlek många intressanta möjligheter för förbättrad analys och det har från början setts som en viktig komponent i det program för speciell satsanalys som byggs. Resonemanget är att en perfekt regelbaserad grammatik för svenska i princip borde innehålla minst den information som extraheras ur ett valenslexikon. Det har förekommit försök att dra nytta av någon av dessa resurser för parsning tidigare, men dessa försök verkar ha avslutats utan att valenslexikon blivit del av implementationerna.[72] Det verkar i sammanhanget relevant att dra en skiljelinje mellan *attributvalenser* och *verbvalenser*, där alltför strikt användning av de senare inte så klart verkar förbättra korrektheten. Den förbättring som faktiskt görs är som visas också blygsam, men välmotiverad. Detta ska dock inte ses som ett slutgiltigt svar på frågan om möjlig språkteknisk användning av dessa utarbetade resurser.

För att skapa ett lexikon som kan användas för att snabbt slå upp syntaktisk valens främst för verb, substantiv och adjektiv, krävdes ett betydande arbete av sortering, omformning och borttagning av redundant information.

---

[72] Det finns förmodligen inte något annat parsningssystem idag som använder valensinformationen som här omarbetats, även om försök lär ha förekommit med data från *Lexin – Svenska ord* (1998).

4.5.1 Valenslexikonet i *Nationalencyklopedins ordbok (NEO)*



1	Uppslagsord	Ordklass	Valens	Kommentar
5658	spela 1/1	verb	& ngt	
5659	spela 1/2	verb	& (ut) ngt, & PRED	
5660	spela 1/3	verb	& in/upp ngt	MOTS, X-lexem omkast
5661	spela 1/4	verb	& (ut) ngt, & ngt (med ngn), & på ngt	
5662	spela 1/5	verb	& på ngt	
5663	spela 1/6	verb	&&	
5664	spela 1/7	verb	& (ngt)	
5665	spela 1/8	verb	& (ADVL)	
5666	spela 2/1	verb	& ngt (ADVL)	
5667	spelbar 1/1	adj.	& (för ngn)	
5668	spelmansstämma	subst.	(på/under/vid) &n	

Figur 37 Ett utdrag från en valenskälla (*NEO*) med olika lemman och lexem visar de ofta många potentiella komplementstyperna för ett verb som *spela*. *&&* betyder intransitiv användning.

Den information som erhållits för användning i denna tillämpning är listor av uppslagsord i grundform tillsammans med beskrivning av syntaktisk valens.[73] Beskrivningen av valenserna i uppslagen har drag av formalitet som liknar reguljära uttryck såsom parenteser för optionalitet som i *”& (ngt)”*. Ampersand står för uppslagsordet (eventuellt böjt). ”/” står för alternativitet som i *”& ngt/ngra”*. De valenslistor som använts har varit sorterade med avseende på lemma och lexem. Det har inneburit en form där vissa formmässigt lika ingångar ofta duplicerats flera gånger i lexikonet, utan att det riktigt är uppenbart för den som läser lexikonet i denna form (utan betydelseangivelse) vilken exakt skillnad som finns mellan de olika versionerna. I det sammanhang här, som denna valensinformation är tänkt att användas, finns ju ingen sådan semantisk distinktion gjord. En grundform med specificerad ordklass är istället tänkt att genom lexikonet ge grundval för olika syntaktiska val. Det hela sker utan statistik om vanlighet för de olika möjliga konstruktionerna och resultatet som förväntas är att möjliga syntaktiska valenser för alla lexem (från eventuellt flera lemman med samma grundform) hämtas från lexikonet.

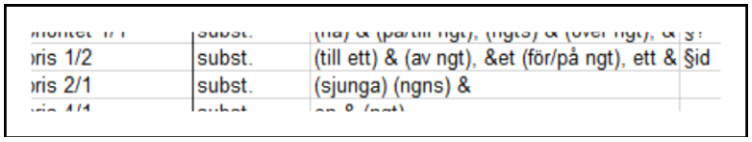
Som kan ses ovan är det vanligt att olika betydelser med samma grundform har samma valens. Detta är dock distinktioner som helt försvinner i utdataformatet

[73] Utgångsinformationen har varit tillhandahållen data från Språkdata genom områdesspecialisterna Sven Göran Malmgren och Maria Toporowska Gronostaj, som arbetat med detta inom NEO. Noga räknat är det som här kallas NEO ett bakgrundsmaterial till denna ordbok och även till senare *Svensk ordbok*.

eftersom det helt enkelt samlar alla möjliga valenser under ett uppslag: *verb_spela*.

& ngt	& (ut) ngt	& ngt (med ngn)
& (ut) ngt	& ngt (med ngn)	& på ngt
& PRED	& på ngt	& på ngt
& in/upp ngt	& (ut) ngt	&&
& (ut) ngt	& ngt (med ngn)	& (ngt)
& ngt (med ngn)	& på ngt	& (ADVL)
& på ngt	& (ut) ngt	& ngt (ADVL)
& (ut) ngt	& ngt (med ngn)	& på ngt
& ngt (med ngn)	& på ngt	& (ut) ngt

Tabell 41 Förskönad (ej komprimerad) version av faktisk utdata från uppslag av *verb_spela*. Som synes förekommer samma valensuttryck ett flertal gånger och många konstruktioner finns implicit inom andra uttryck. I aktuell version sker knappt någon sammansmältning av dessa.



Figur 38 Informationen som finns i NEO innebär oftast specifikationer för komplementsled men det finns också exempel på vilka konstruktioner som kan föregå substantiv som för *pris* här. En information som *sjunga någons pris* används dock inte här.

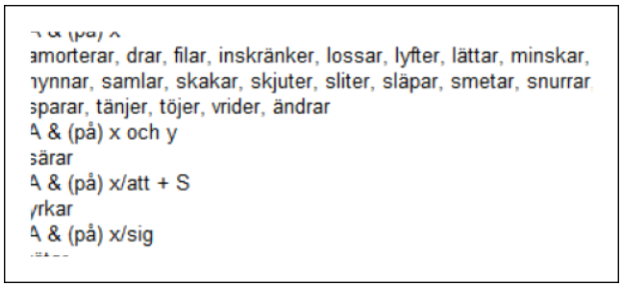
Eftersom det inte utan vidare går att specificera vilken betydelse av potentiellt flera som är den riktiga i ett visst sammanhang i texten, innebär ett så rikt utbud som det för *spela* en begränsad hjälp. Andra uttryck har mindre konstruktionspotential och NEO ger för dessa bättre möjligheter till att rätt knyta objekt m.m. till valensordet.

I det ursprungliga formatet, se Figur 37, finns drygt 19 400 rader av uppslag där dock uppslagen slås ihop när de behandlar ord med samma ordklass och gemensam grundform. Det resulterade i ett verbvalenslexikon med färre än 7 000 ingångar, för substantiv ett liknande antal ingångar och ett för adjektiv ca 1 800 ingångar.[74]

[74] Därutöver fanns i lexikonet ett femtontal ingångar för övriga ordklasser som adverb: *ont – & om ngt* och substantivisk förkortning: *VM – & (i/på ngt).*

4.5.2 Valenslexikonet i *Lexin – Svenska ord*

I den andra källa som undersökts och använts, *Lexin – Svenska ord,* finns en annorlunda typ av information än i NEO. Den finaste fördelen med detta lexikon är att det innehåller information om subjekt förutom komplementsdelen, och att animathet, närmare bestämt människa/ting, finns kodat. A och B betyder personer (eller ibland t.ex. en organisation som kan agera som en människa) medan x och y står för ting. Ampersand står även här för uppslagsordet. I *Lexin – Svenska ord* ser ett uppslag t.ex. ut enligt följande: *Kurar ihop sig*: *A &*. Att både subjekt och komplementsdel finns med leder till att en riktig mening formas genom att byta ut ampersand mot uppslagsordet. (Det är förmodligen därför som detta skrivits i presens, i en inlärarinriktad ordlista som detta är.)



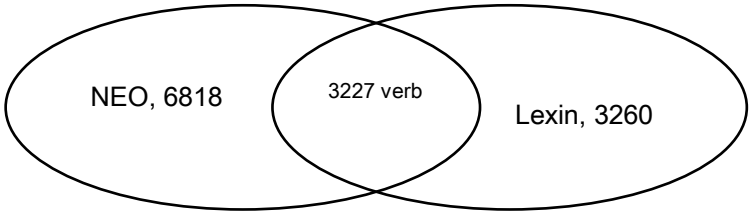
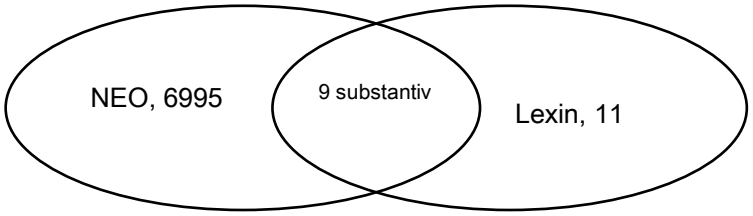
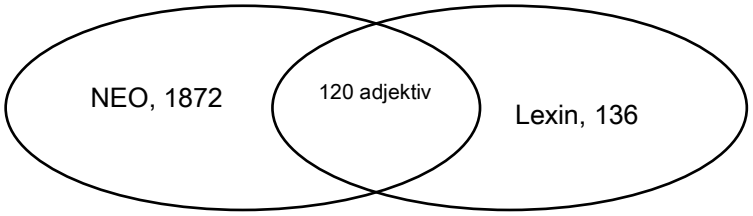
Figur 39 Den variant av *Lexins* valenslexikon som använts är sorterad på valenser.

En motsvarande förbehandling genomfördes för Lexins valenslexikon. Eftersom uppslagsorden i detta valenslexikon står i presens gjordes en omformning till grundform till stor del manuellt.[75] (Se avsnitt om grundformsfunktionalitet nedan.) En skillnad mellan *Lexin* och *NEO* är att uppslagen varierar lite i de två källorna när det gäller huruvida reflexiva partiklar (och olika småord) finns med som valensdel eller som namn på uppslag. T.ex. finns i NEO uppslaget *Kura* med valensvärdet *& (ihop) (sig) (ADVL)* medan det i *Lexin – Svenska ord* alltså finns ett uppslag som heter *Kurar ihop sig* med valensvärdet *A &*. I uppslagsmekanismen som byggts sker uppslag på ett enda grafiskt ord. Detta har föranlett att alla verb med reflexiv, partikel etc. har omformats så att enbart verbet är uppslagsnyckel. Genom att slå ihop ingångar på det sättet sjunker speciellt antalet verbvalensingångar (värdena blir däremot uppdelade på fler valensfall).

[75] Omkring 65 % av ingångarna slutar på *”-er”* eller *”-ar”* och fick grundform när denna ändelse byttes mot *”-a”*. Därutöver finns olika oregelbundna verb, verb med s-form och uppslag som består av mer än ett verb i presens.

4.5.3 En jämförelse mellan Lexin – *Svenska ord* och NEO

Tabell 42 och Figur 40 är jämförelser av antalet extraherade lexikoningångar i NEO och *Lexin – Svenska ord*. Alla siffror rörande antalet ingångar gäller i ett läge där samtliga uppslag med samma grundform och ordklass slagits samman till en enda ingång eftersom det är okänt vilken variant som faktiskt är den påträffade i en text. Efter att denna omformning skett är samlingarna i följande slutliga storlekar.

Verb finns det gott om i båda resurserna. Även om de flesta av *Lexins* också finns i NEO så är valensinformationen annorlunda.	
NEO har ungefär lika många valensingångar för *substantiv* som för verb.	
Adjektiv behandlas också. NEO har större täckning.	

Figur 40 En slutsats att dra från ovanstående värden är att de flesta ingångarna i *Lexin* finns i NEO. Inte desto mindre är informationen i *Lexin* unik bl.a. genom subjektsbeskrivning.[76]

[76] Flera personer har reagerat på det här uppvisade lilla antalet ingångar för vissa ordklasser i *Lexin*. Detta var emellertid det faktiska antalet i den samling med substantivvalenser som hanterades i detta arbete.

	NEO	*Lexin*
Antal verb	Ca 6800	Ca 3200
Antal substantiv	Ca 6900	Ca 10
Antal adjektiv	Ca 1800	Ca 130
Subjektsinformation i valensen	Nej, generellt inte	Ja – mänsklig/inanimat klargörs
Komplementsinformation i valensen	Ja, med en sorts animathetsinformation: ”NGT/NGN” etc.	Ja. Animathet: A/B/C är mänskliga. x/y/z är ting.

Tabell 42 Jämförelse av lexikonstorlek m.m. visar att materialet som ligger till grund för NEO är klart större än det för *Lexin*. Siffrorna gäller slutformen och är ungefärliga.

4.5.4 Grundformsfunktionalitet

Grundformsfunktionaliteten har en uppgift: att ta en eventuellt böjd form av ett verb, substantiv eller adjektiv och returnera grundformen för att kunna undersöka eventuell valens. Funktionen utgår från ett ord tillsammans med taggad ordklass och ska leverera grundform. (Det är egentligen tre olika funktioner – för substantiv, verb och adjektiv.)[77] Grundformslexikonet är anpassat till samlingen valensdataingångar och ska precis täcka de lexem/lemman som finns i valenslexikonet. Metoden blev därför att utgå från valenslexikonet[78]. För att ge grundformer skapades en funktion som tar bort suffix och lägger till vissa typiska grundformsändelser tills ordet förvandlats till en grundform i valenslexikonet. Detta kunde göras eftersom substantiv, verb och adjektiv oftast böjs på ett fåtal återkommande sätt. Sedan användes denna omformningsmetod med *Lexin*-lexikonet (inte *valenslexikonet*, utan den fulla versionen). För varje uppslag som fanns i det byggda valenslexikonet hämtades alla former och dessa undersöktes. Om omformningsfunktionen lyckades ge grundformen för en böjningsform (t.ex. härleda *framgång* av *framgångarnas*) accepterades det. Kodexempel 8 är ett exempel på en ingång i *Lexin*-lexikonet (som också omformades för ändamålet).

[77] Denna funktionalitet byggdes före den fria resursen *SALDO* (Borin, Forsberg och Lönngren 2008) med grundformsfunktionalitet, gjordes tillgänglig.

[78] Ett första försök var att skapa ett lexikon med hjälp av SUC där varje löpord presenteras tillsammans med sin grundform. Detta genomfördes, men en svaghet i den aktuella tekniska lösningen är att detta lexikon alltså bara täcker precis de ord som finns i korpusen. En idé var att använda hela *Lexins* lexikon (det vanliga lexikonet, inte valenslexikonet!). Detta skulle lett till ett mycket stort lexikon och det sågs som en praktisk svaghet eftersom inladdningstiden helst inte skulle öka alltför mycket.

Kodexempel 8 `("verb_absorberar","absorberade absorberat absorbera");`

För *absorberar* tas alltså *absorberade*, *absorberat* och *absorbera*. Dessa tre ord undersöks sedan ett efter ett och det klargörs om de genom avkapningar och tillägg kunde ge grundformen. Det innebar att grundformsfunktionen förbättrades stegvis när listan över överblivna ord gicks igenom. Detta enkla arbetssätt föll relativt väl ut och de oregelbundna ord som inte lyckats ledas till grundformen, eller som leder till fel grundform, placerades i ett undantagslexikon.

4.5.5 Hur ofta är valensinformation till nytta för attributbestämning?

Det är i programmet, som beskrivits, en fråga om en stor samling av valensinformation, främst bestående av prepositioner som inleder PP-attribut till substantiv, adjektiv och particip. I praktisk användning löser informationen en *PP-attachment*-fråga som i Ex 61 och sammanfogar segment. Med en ’träff’ avses här att en preposition i en följande prepositionsfras finns i listan av attributiva prepositioner hos föregående NP-huvudord i grundform. Detta innebär att bara verkligt betydelsefulla sammanfogningar på dessa grunder räknats, och inte t.ex. sådana där en matchning sker inom ett segment som redan sammanfogats till samma chunk på andra grunder. När valensinformation från lexikonen anger att attributrelation gäller knyts delar samman, som t.ex. *tyngd – av* i Ex 61.

Ex 61 […] hennes verk är tyngda av existentiellt allvar, av ceremoniell högtidlighet. (cc03e-009)

En manuell undersökning av framslumpade s-enheter från SUC gav nedanstående frekvensresultat avseende antal sammanfogningar av chunkar genom valensmatchning från en föregående strukturs huvudord (ifall denna är PP-formad avses sista ordet i segmentet) mot inledande preposition i följande PP. Det antal träffar och faktisk användning vid analys som anges gäller den analysform som här utförs, dvs. huvudsatsanalys som begagnar sig av sammanfogning på grundval av uteslutning. Frekvensen beror också i hög grad på hur ofta segment sammanfogas på andra grunder i syntaxanalysatorn – t.ex.genom identifikation av bisats, då valenslänkningen blir överflödig just där.

---

203 av **10 000,** dvs. **ca 2 %** slumpvis analyserade s-enheter från träningsmängden innehöll länkning av segment genom attribut (substantiv, adjektiv eller particip till preposition) från NEO-databasen

---

Frekvensuppskattning 16 Frekvens för användningen av attributvalenser från databasen som ligger till grund för NEO.

Det är en inte oviktig poäng att den relativt lilla nytta som attributvalenserna för med sig skulle öka om systemet även gjorde analys på underordnade satsnivåer. Underordnade satser sammanfogas nu oftast heuristiskt utan ingående analys.

5 Automatisk textvariation samt automatgenerering av besvarade frågor från text

I detta kapitel beskrivs direkt användning av text som ges analys enligt föregående kapitel. De två prototyperna som presenteras är tekniskt sett del av analysprogrammet och gynnas därmed direkt av den kontinuerliga förbättringen i korrekthet som analysprogrammet möjliggör. Prototyperna använder ordklasstaggning och får analysresultatet internt i det beskrivna XML-formatet. Båda tillämpningarna är byggda med teknik som kräver minst just det beskrivna syntaktiskt funktionella formatet på huvudsatsnivå. Den andra tillämpningen saknar helt föregångare även prototypmässigt, åtminstone för svenska.

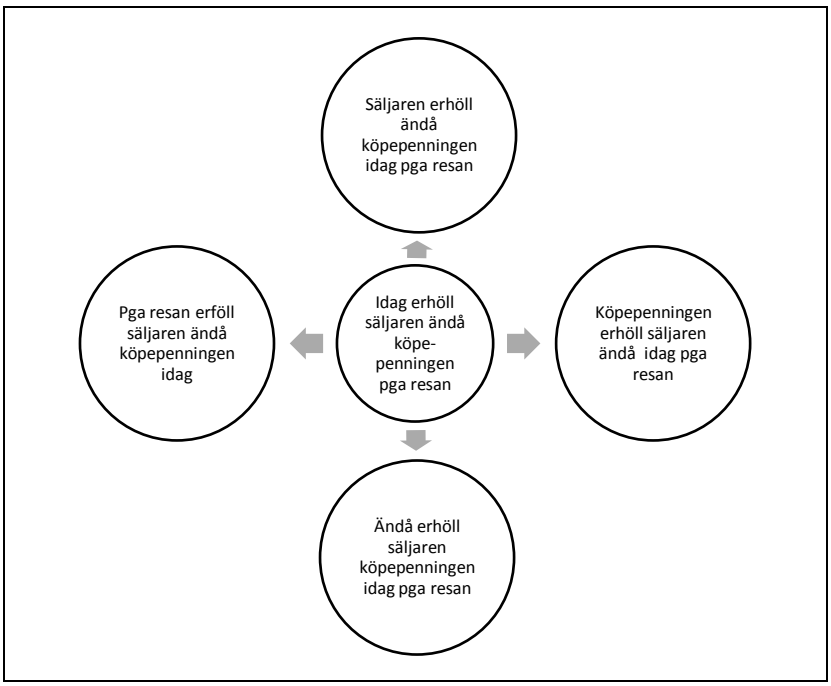
Struktur i kapitel 5, Automatisk textvariation samt automatgenerering av besvarade frågor från text

I kapiteldel 5.1 *Automatisk variation av svensk text genom spetsställning*, behandlas automatisk omformulering av svensk text från ett teoretiskt och tillämpningsmässigt perspektiv. Spetsställningsparafrasering har varit ett omforskat område teoretiskt men det finns bara ett fåtal tillämpningar. Programmet är en prototyp i editorliknande miljö. Det berörs dock hur funktionaliteten skulle kunna transformera syntax i svensk text slump- eller regelmässigt för att studera den effekt detta får för textkoherens, naturlighet och sanningsvillkorsbevarande.

I kapiteldel 5.2 *Automatisk generering av besvarade frågor från text*, redovisas ett program av typen *natural language query system*. Genom att med analysen som grund producera *hv*-frågor är strategin att låta systemet generera möjliga frågor till en text, webbsida etc. som öppnas. Det innebär att frågor som ställs på naturligt språk kan begränsas till de som faktiskt har genererats och, om analysen är korrekt, besvaras per definition. Metoden innebär ett slags alternativ till en sökmetodik som använder mönstermatchning och som riskerar att leverera bara ’bästa matchning’ eller felaktiga svar. Programmet besvarar frågor som skapats till en valfri indatatext eller den nedladdade textdatabasen till Wikipedia. Mer avancerad utvinning av propositioner görs för närvarande inte, delvis därför att *precision*-värdet (rätt svar givet frågorna) då riskerar att bli lidande. Inte heller löses anaforisk referens i programmet utan svaret på en fråga ges genom att scrolla till meningen som frågan genererats från, vilken markeras så att text ovanför också visas och därmed möjliggör för användaren själv att lösa anaforer etc.

5.1 Automatisk variation av svensk text genom spetsställning

Svenska är såsom satsschemat klargör ett språk med en ordföljd mittemellan fri och fixerad (*moderately fixed*). Satsschemat kan pedagogiskt användas för att visa hur fundamentering av primära och vissa utbrutna satsled ger grammatiskt resultat med bibehållna sanningsvärden. Datorimplementationer för att åstadkomma sådana parafraser av svensk text, utgående från schemat eller ej, har däremot hittills varit få och typen finns inte veterligen i något kommersiellt eller publikt tillgängligt system.



Figur 41 En spetsställning av de andra primära obegränsade leden i enkel adverbialinledd huvudsats ger fyra andra möjliga parafraser med samma betydelse. (Även adverbialet *ändå* spetsställs här.) Ofta ses en form där subjekt finns i fundamentet som grundformen i svenska varifrån spetsställningar utgår. Här kommer även den att kallas för en sorts spetsställning (av subjektet).

Möjligheten till spetsställning av de flesta obegränsade led, enligt Figur 41, i svenska öppnar för några intressanta frågeställningar. I detta kapitel frågas i vilken mån spetsställningsvarianter är *grammatiskt giltiga, naturliga och sanningsvillkorsbevarande*. Frågan är om en mekanisk funktionalitet för spetsställning kan genomföras i svensk text på ett okomplicerat sätt. I bakgrunden för denna frågeställning finns förhoppningen om att kunna skapa innehållsmässiga ekviva-

lenser, dvs. en sanningsvillkorsbevarande process (*salva veritate*).[79] Vad som däremot förändras med spetsställningen av led i en sats är dess *informationsstruktur*.

Denna kapiteldel innehåller exempel på de vanligaste transformationstyperna som förekommer i litteratur om svenska. Ett program för användarinitierad spetsställning av primära led i editorsituation, byggt på den beskrivna syntaxanalysen, presenteras.

Struktur i kapiteldelen 5.1, Automatisk variation av svensk text genom spetsställning

I avsnitt 5.1.1 visas exempel från tre olika försök till omformulering av svensk text.

Avsnitt 5.1.2 svarar på frågan om i vilken utsträckning olika satsled förekommer spetsställda i svensk text rent kvantitativt, enligt litteraturen. En närliggande fråga som tas upp är den om genomsnittlig fundamentlängd, vilket brukar tilldelas en speciell betydelse i textstilanalys.

Avsnitt 5.1.3 behandlar den grundläggande frågan om vad spetsställning egentligen har för språklig funktion i svenska och något om hur denna fråga hittills besvarats i litteraturen.

Avsnitt 5.1.4 beskriver de grundläggande begränsningar som gäller för spetsställning av primära led i svenska. Avsnittet beskriver förbjudna flyttningar med avseende på grammatikalitet och bevarande av en huvudsats innehåll.

Avsnitt 5.1.5 är en redogörelse av en prototypimplementation för svensk text byggd på den syntaxanalys som här beskrivits i föregående kapitel. Denna har byggts som en editor och tillåter användarinitierad omformulering genom gränssnittet. En kortfattad beskrivning av denna finns i Wilhelmsson (2008).

Avsnitt 5.1.6 är ett avslutande stycke om spetsställning och tar upp några andra möjligheter som finns för syntaktiska meningsbevarande transformationer på främst huvudsatsnivå.

[79] Termen *Salva Veritate* för denna egenskap hos transformationer användes av Gottfried Leibniz.

5.1.1 Parafrasprogram för svenska

Fältet *paraphrasing* inom språkteknologin har tilltagit internationellt under senare år, men det svarar bara delvis mot uppgiften här: att formulera samma innehåll på flera olika sätt. Bl.a. inbegriper området forskningsuppgiften att avgöra när två diskurser eller textmeningar avhandlar samma fakta, en metod som kan tillämpas för att jämföra två tidningsartiklar på detta sätt. Ett exempel för engelska är *Microsoft Research Paraphrase Corpus*[80] som tillhandahåller 5 800 textmeningspar som av två annoterare bedömts som parafraser eller ej. I det fallet är parafrasbegreppet väldigt löst definierat och exemplifieras med Ex 62. Med andra ord handlar det inte om generering av parafraser utan om bedömning av huruvida enheter är 'närparafraser'.

Ex 62 The genome of the fungal pathogen that causes Sudden Oak Death has been sequenced by US scientist ↔ Researchers announced Thursday they've completed the genetic blueprint of the blight-causing culprit responsible for sudden oak death

För svenska finns några försök att som i detta arbete *generera* parafraser utifrån given text. Dessa ligger närmare uppgiften att sanningsvillkorsbevarande skapa textvarianter. Att automatisk syntaktisk omformulering i form av spetsställning hör ihop med satsschemat när det gäller svenska är ett utgångsläge i *Topikalisering som skrivstöd. En implementering med satsschema* (Lindberg och Svensson 1992). Detta arbete innefattade analys av utvalda meningar som analyserades med hjälp av *MorP* (Källgren 1992) och därefter omvandlades med hjälp av ett redovisat Prolog-program. Genomgången av spetsställningar i arbetet verkar vara ett pionjärarbete och visar på begränsningar i parafrasmängden på olika grammatiska grunder. Ett exempel på automatisk omformulering av svensk text utan lika tydlig koppling till satsschemat finns i ett arbete (Pascoe och Ullner 2006) som undersöker automatiskt diatesskifte utgående från text syntaxanalyserad av *CassSwe* (Kokkinakis 1998). Textvariation kan också ske på lexikal nivå, genom synonymutbyte, vilket undersöktes i ett experiment av Rosell (2005), *Variator*, med hjälp av *Folkets synonymordlista* (Kann och Rosell 2005)[81] som tillhandahåller synonympar tillsammans med ett numeriskt värde på en skala upp till 5,0 och som enligt en användargrundad bedömning anger hur bra synonymer de antas vara, t.ex. *slå – aga: 5,0*; *slå – banka: 4,2*. (Se även avsnitt 5.2.)

[80] Denna korpus kan laddas ned från http://research.microsoft.com/apps/dp/dl/downloads.aspx (senast bevistad 18 augusti 2009).
[81] Denna ordlista kan laddas ned från http://lexin2.nada.kth.se/synlex.html (senast bevistad 22 oktober 2009).

Implementationerna som rör *syntaktisk* omformulering enligt Tabell 43, spetsställning och diatesskifte, är inte helt utformade för fri text, speciellt spetsställningsprogrammet (i mitten) hanterade handplockade parsade meningar.

Textklustring, åtminstonde som det beskrivs här, utnyttjar sig av den vektorrumsmodell, som används allmänt inom området. ⇳ *Textklustring, åtminstonde som det skildras här, exploaterar sig av den vektorrumsmodell, som brukas offentligt inom området.* ⇳ *Textklustring, åtminstonde som det beskrivs här, utnyttjar sig av den vektorrumsmodell, som brukas allmänt inom området.*	*Vanligtvis brukade vi studera blommor här på mornarna.* ⇳ *Blommor brukade vi vanligtvis studera här på mornarna.* ⇳ *Här brukade vi vanligtvis studera blommor på mornarna.* ⇳ *På mornarna brukade vi vanligtvis studera blommor här.*	*Bron som än idag leder över från gamla stan till Hradcany byggdes på 1300-talet.* ⇳ *Man byggde bron som än idag leder över från gamla stan till Hradcany på 1300-talet.* *Lådor med starköl och kassar med vodka bärs i riktning mot hamnen.* ⇳ *Man bär lådor med starköl och kassar med vodka i riktning mot hamnen.*
Exempel på synonymiutbyte från Rosell (2005)	Exempel på spetsställning från Lindberg och Svensson (1992)	Exempel på diatesskifte från Pascoe och Ullner (2006)

Tabell 43 Textvariation av svensk text har implementerats genom lexikal och syntaktisk variation. Synonymiutbyte, t.v., ställer höga krav på synonymivärde för att inte ge ett oklart eller komiskt uttryck. De två modellimplementationer som rör *syntaktiska transformationer* ställer istället krav på korrekt parsning. Den bitvis satsschemagrundade ansatsen programmerad i Prolog, i mitten, och undersökningen om huruvida det är möjligt att skapa aktiva satser av passiva, till höger, representeras här med lyckade exempel.

5.1.2 Hur ofta är olika ledslag spetsställda i svenska?

Den föreliggande uppgiften behandlar spetsställningstransformation och i princip fri, okänd text. Den första frågan som här har ställts är hur vanligt fundamentplacering av olika led är naturligt i svensk text, rent kvantitativt. Sådana beräkningar finns på flera håll i litteraturen. I Westman (1974) anges siffrorna i Tabell 44 för olika spetsställda led (där subjekt också ses som spetsställt) i

svensk tidningstext. Tabell 45 visar resultat från en mindre undersökning (200 meningar från böcker och tidningstext) i Lindberg och Svensson (1992), men där subjekt inte räknas, dvs. subjekt i fundamentet räknas som den kanoniska formen.

Subjekt	Adverbial	Objekt	Predikativ	Övriga
63 %	29,6 %	4,2 %	2,2 %	1,1 %

Tabell 44 Spetsställda adverbial har följande interna fördelning. Objekt sägs här utgöra en relativt stor andel p.g.a. att anföringsled ses som objekt.

PP	Satsadv	Adv-bisatser	Objekt (ord)	Adj-fras	Infinit sats	Obj-bisatser	Utbr bi-satsled
75 %	13 %	4 %	3,5 %	2 %	1 %	1 %	0,5 %

Tabell 45 De ovanstående är fundamentinnehåll där subjekt ej räknas som spetsställning.

5.1.3 Vilken funktion fyller spetsställning av satsled i svenska?

De ovanstående mätvärdena visar en viss variation beträffande förekomsten av fundamenterade led. Men vad betyder då fundamentering av ett led – vad är det som uttrycks med en viss spetsställning (och som inte kan göras på riktigt samma sätt i ett språk med mer fixerad ordföljd, som engelska)? I Lindberg och Svensson (1992) nämns följande *orsaker* till spetsställning i svenska: emfas, kontrast, referens till något som sagts omedelbart innan samt variation i satsbyggnad. Det måste konstateras att de olika orsakerna som nämns har att göra med de krafter som brukar beskrivas som *textbindning*. Undantaget är då 'variation i satsbyggnad'. Frågan är hur öppen för variation satsbyggnaden verkligen kan vara för textmeningar i kontext.

I Engdahl (1999) beskrivs spetsställningens grammatiska villkor i svenska, och det undersöks speciellt hur spetsställning får göras med avseende på den *pragmatiska* funktionen. Det klargörs att spetsställning kan ske med de flesta ledslag (med individuella och kontextmässiga undantag). Speciellt för svenska är att spetsställning kan ske med ett led som utgör obetonad (icke-kontrasterande) topik som i Ex 63. [82]

Ex 63 Hon hade just fått smågrisar, nio stycken. Dom tyckte Kerstin om att titta på _.

[82] I Engdahl (1999) används *topik* som den förankrande utgångspunkten och *rhema* som den nya informationen, t.ex. den information i en sats som direkt besvarar en fråga.

Medan andra språk uppvisar begränsningar ifråga om spetsställning av led som utgör obetonad topik, finns för svenska ingen klar begränsning. Med andra ord verkar spetsställning som transformation ha förutsättningar att *kunna* fungera mer obehindrat än på andra språk och det är svårt att tilldela den en helt klar pragmatisk funktionsroll. (För engelska är sambandet däremot tydligare: *topik* framflyttas – funktionen kallas *topicalization*).

5.1.4 Vilka begränsningar finns för spetsställningar i svenska?

I fråga om begränsningar i den mekaniska spetsställningsprocessen kan dessa delas in i dels konkreta begränsningar rörande grammatikalitet och dels begränsningar rörande betydelsebevarandet.

5.1.4.1 Grammatiska begränsningar

Det finns några tydliga exempel på primära led som ej kan fundamenteras. (Denna beskrivning rör alltså *primära* led, attribut kan generellt ej spetsställas och satsflätor etc. hanteras ej här). Spetsställning är för det första inte grammatiskt giltig för s.k. obetonade talaktsadverbial, nedan huvudsakligen hämtade från Lindberg och Svensson (1992), även om somliga förekommer i lätt avvikande poetisk text osv.

Ex 64 *ju, väl, icke, verkligen, också, ej, nämligen, väl*

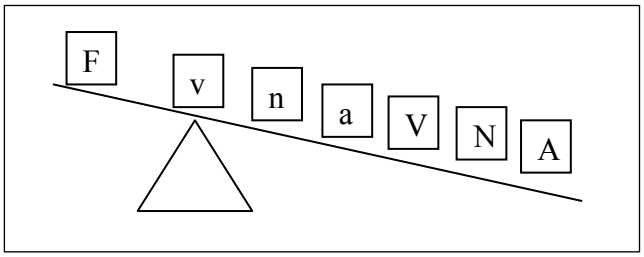
Härtill kommer flyttningsbegränsning för *heller* som eventuellt tolkas som eget fristående led i konstruktion tillsammans med *inte* (så kan det ju tolkas ifall *inte* spetsställs: *Inte sprang de heller*).

Andra indirekta begränsningar gäller formen hos sekvensen som blir resultatet efter finitet och har t.ex. beskrivits av Andersson (1974). Ex 65 visar hur spetsställning av subjekt, *jag*, kräver att typerna av adverbial placeras i en speciell ordning när *I hela tjugo år* ska flyttas från fundamentet för att ge plats åt subjektet. Andersson påpekar att den andra versionen ändå inte är ogrammatisk, vilket egentligen borde förpassa detta exempel till de med semantiska begränsningar i nästa avsnitt.[83]

Ex 65 *I hela tjugo år hade jag semester i tre veckor* ←//→
Jag hade semester i hela tjugo år i tre veckor

[83] Andra av Anderssons exempel verkar tydligare bryta grammatikaliteten genom att inte ha adverbialen i den hierarkiordning han beskriver:
? Skidat i två timmar fyra gånger har jag gjort på tre veckor.

Mer allmänna krafter som påverkar *naturligheten* men inte alltid omöjliggör konstruktioner finns också. Dessa regler leder t.ex. till att somliga former blir vanligare i praktiken. I litteraturen tas upp begränsningar i den faktiska variationen som både kan kallas grammatiska och semantiska genom *tyngdlagen* (enligt Figur 42), samt att känd information gärna placeras framför ny och att informationen ska organiseras ”så att den lättast kan förstås och möjligheter till feltolkning elimineras” (Jörgensen och Svensson 1986).



Figur 42 En av de generella satsinterna krafter som inverkar på ordföljden är den s.k. *viktprincipen*. Denna innebär att tunga, långa led tenderar att placeras längre bak i satsbyggnaden. I huvudsatsperspektivet innebär det att en högertyngd satsbyggnad har företräde enligt gungan på bilden, närmast hämtad från Melin och Lange (1986), s. 71. I denna ’naturliga’ situation befinner sig huvudsatsens finita verb vänster om mittpunkten i satsen och antalet ord är fler till vänster än till höger (ledbokstäver representerar konstituenter snarare än positioner).

Lång fundamentlängd, s.k. vänstertyngdhet, är således en egenskap som enligt principen inte gynnas men förekommer bland SUCs mest skriftbetonade språk.[84] Måtten fundamentinnehåll och även *fundamentlängd* har framträdande platser som ytteknisk textstilmarkör, t.ex. i Landqvist (2000), se Tabell 46. Här antas långa genomsnittliga fundamentlängder generellt vara tecken på ett lite mer stelt och formellt språkbruk.

Kåseri	Sport	Ledare	Kultur	Totalt
2,8	3,0	3,9	3,4	3,2

Tabell 46 Fundamentlängd i medeltal ord redovisas här från en undersökning av Melin och Lange (1986), s. 169.

[84] Ett exempel i SUC, jb04-057: *Även om vi för egen del har svårt att rätt uppskatta det rationella i en strategi i utbildningsfrågan som innebär, att man når en förändring av undervisningsmetoderna endast genom organisatoriska omläggningar som medvetet sätter lärarna i en pedagogisk tvångssituation, i vilken det inte längre går att upprätthålla det gamla undervisningsmönstret, kan det knappast förnekas att grundskolereformen kommit att spela just den rollen på högstadiet.*

5.1.4.2 Semantiska begränsningar

En annan begränsning för helt blint mekaniska transformationer rör den svenska semantiken snarare än grammatikaliteten hos ett parafraspar. I vilken mån kan sanningsvillkorsbevarande garanteras när led spetsställs i enlighet med satsschemat? Ifråga om enskilda ord är satsadverbialet *nog* ett speciellt icke spetsställbart led i Ex 66.

Ex 66 *Nog funkade det ←//→ Det funkade nog.*

Ett annat undantag gäller *kvantifierare som subjekt*, vilka interagerar med satsadverbial så att dessa får en betydelseskiljande funktion, som i Ex 67 från Jörgensen och Svensson (1986).

Ex 67 *Fortfarande har inte någon anlänt ←//→ Fortfarande har någon inte anlänt*

Ett tredje konkret exempel på hur den bibehållna betydelsen kan äventyras genom spetsställning är den ambiguitet som uppstår vid kombinationer av verb och nominala ledpar som låter placeringen avgöra vilket led som är subjekt respektive objekt som i Ex 68.

Ex 68 *Ekonomin påverkade börsen ←//→ Börsen påverkade ekonomin.*

Ex 68 verkar visa den begränsningstyp som kräver flest konkreta undantagsregler för en implementerad funktionalitet. Implementationstekniskt är de grammatiska och semantiska begränsningarna lika såtillvida att de kan formuleras som undantagsregler byggda på ordlistningar.

Ytterligare begränsningar rörande möjligheten att syntaktiskt omformulera med bevarade sanningsvillkor finns i Jespersens (1924) grundläggande uppdelning mellan *fixed expressions* och *formulas*. De flerordskonstituenter som får betydelse på ett *icke-kompositionellt* sätt på svenska inbegriper *idiom* (talesätt) och *ordspråk,* där de senare till skillnad från de förra inte låter sig omformas syntaktiskt med sammanhanget.

5.1.5 Implementation av användarinitierad spetsställningsparafras i editormiljö

Den prototyp som byggts i detta arbete är utdatamässigt inte olik den av Lindberg och Svensson (1992) men applicerad i en editeringssituation med grafiskt gränssnitt och därmed riktad mot fri (okänd) svensk text. Detta innebär att ordklasstaggning och syntaxanalys sker vid inskrivandet av varje textmening. Det är ett grundläggande antagande att den syntaxanalys som krävs behöver ha identifierat fullständiga led, dvs. nominala och adverbiella strukturer med hela sträckningar, inklusive efterställda attribut och så vidare, eftersom det allmänt är

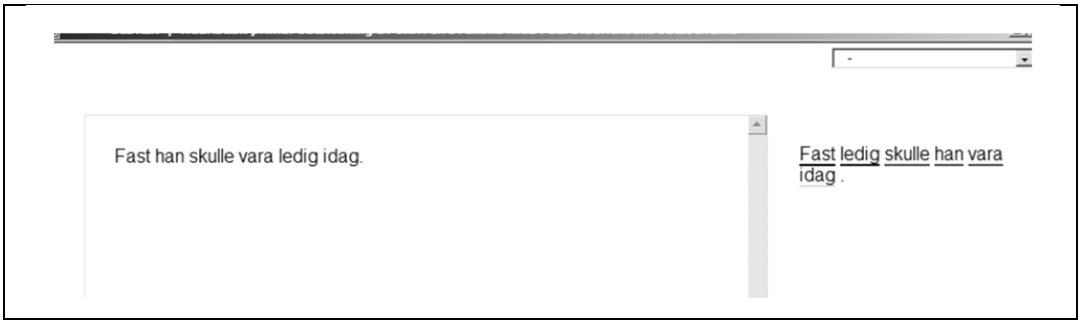
dessa segment som hanteras vid grammatiska omflyttningar. Lindberg och Svensson (1992) beskrev visserligen också funktioner av utbrytning av underordande satsnivåer, men det gällde över huvud taget iordningställda analyser.

Implementationen är relativt försiktig i aktuellt utförande och gäller alltså möjlighet till spetsställning av fullständiga enstaka primära led – dvs. utdata från programmet som beskrivits i föregående kapitel. Satsflätor produceras alltså inte, och inte heller spetsställning av hel verbfras (dvs. innehåll från mer än en satsposition), ett spetsställningsfenomen som förekommer relativt frekvent, enligt exemplen från SUC i Ex 69.

Ex 69 a) Men tala om sjukhus vågade ingen, sådant tal skulle inte uppskattas. (kk48-006)

b) Avsade sig gjorde också föreningens kassör Ingvar Johansson, Sala, efter 40 år. (af06d-009)

Programmet (se Figur 43) som låter användaren skapa parafraser genom att klicka på satsled som spetsställs bygger på omedelbar (textmeningsvis) ordklasstaggning och syntaxanalys som visas för den som skriver. Själva spetsställningsproceduren innebär att den interna representationen och XML-formatet transformeras. Den interna proceduren för spetsställning är relativt okomplicerad och sker genom att ett visst led placeras främst (eller efter ett förfält) i en räcka där det tidigare fundamentledet lyfts bort. En efterföljande procedur placerar därefter in det tidigare fundamenterade ledet i en acceptabel position, enligt satsschemat, senare i satsen. Till skillnad från den mer avancerade process som beskrivs i nästa kapiteldel hanterar denna process bara textmeningar som består av en huvudsats, dvs. ej satskedjor eller textmeningar med samordnade primära verbfraser. Anledningen är att det skulle kräva en hel del extra tid och arbete. Ansatsen kan beskrivas som försiktig också såtillvida att den inte möjliggör spetsställning för befarade ofullständiga eller felaktiga analyser.



Figur 43 Gränssnittet för automatisk spetsställning består av en editorvy, t.v., och en analysvy, t.h. I analysen ges den aktuella huvudsatsanalysen färgkodning med understrykning. Om ett led i analysvyn klickas spetsställs det där. Om meningen i analysvyn efter spetsställning högerklickas ersätter den ursprungsversionen t.v. Det går även att få se den underliggande syntaxanalysen i gränssnittet.

I programmet är det lätt att bygga in begränsningar när det gäller spetsställning av olika led. Detta arbete har dock inte ägnats någon större uppmärksamhet hittills. Uppenbara begränsningar som spetsställning av primära begränsade led har dock förhindrats.

Syftet med att bygga en användarstyrd spetsställningsfunktion är, som ofta i editorsammanhang, att låta användaren slutgiltigt avgöra lämpligheten hos en viss parafras. Möjligheten att däremot på detta sätt omforma text utan att involvera en användare öppnar också för experiment som rör spetsställningens funktion. Det vore t.ex. möjligt att skapa texter med konsekvent subjektsspetsställning, slumpmässig sådan, eller på annat sätt förändra texten och studera vad denna omformulering får för effekt på läsbarhet, textbindning och liknande.

Det är inte helt uppenbart hur spetsställningsfunktionalitet i användning kan utvärderas numeriskt annat än med användarbedömningar. Denna implementation bygger på samma kodbas som syntaxanalysen och är en del av samma program. Det innebär att förbättring av parsningen direkt gynnar denna applikation, och de fel som dyker upp har nästan alltid sitt upphov i syntaxanalysen eller ordklasstaggningen. Tiden har inte tillåtit några användarstudier.

5.1.6 Konsekvenser av spetsställning

Spetsställning är som visats en mycket vanlig strukturvariation i svenska. Från ett logikperspektiv kan spetsställningen uttryckas som en formel från en funktionellt grammatiskt analyserad sats som förutsäger grammatikalitet och (samma) sanningsvärde hos spetsställningsparafraser som görs med satsschemat. Denna möjlighet leder teoretiskt till att möjliga formationer av samma ord till samma innehåll kan bli stort. Antalet permutationer, P av en mängd med n element (här: primära satsled) utan satsschemats begränsningar skulle kunna beskrivas som $P = n!$. Det innebär att en sats med sex led motsvarar $P(6,6)$ och ger 720 permutationer. Men med satsschemats begränsningar är antalet möjliga spetsställningspermutationer istället helt enkelt lika med antalet spetsställbara led minus möjliga grammatiska/semantiska begränsningar (enligt föregående avsnitt).

Spetsställning av primära led är dock inte den enda sanningsbevarande transformationen i svenska. Andra vanliga meningsbevarande omflyttningar i svenska inkluderar de i Tabell 47 som delvis har hämtats från Holm och Larsson (1980) och Jörgensen och Svensson (1986).

Här ska naturligtvis även konstateras att de olika strukturvariationerna kan *kombineras*, ett faktum som leder till en kraftig möjlig variation, även med sats-

schemats begräsningar. Med tanke på dessa möjligheter, där Tabell 47 inte heller är en uttömmande uppställning, exploderar antalet formuleringsmöjligheter i många fall. Med undantag för verbfras i fundamentposition och rektionsframflyttning så rör omflyttningarna i Tabell 47 enkla fullständiga led och skulle därmed kunna implementeras från samma analysnivå som här.

Syntaktisk förändring	Exempel
NA-rockad, främst lätta adverbial	Se avsnitt 3.6.1
Negerade objekt (till mittfält)	*Hon har inte sett något* *Hon har ingenting sett*
Konstruktion med formellt och egentligt subjekt	*En bil står parkerad på gatan* *Det står en bil parkerad på gatan*
Verb(-fras) i fundamentposition	*Han spelade fotboll* *Spelade fotboll gjorde han*
Passivisering/aktivering	*Mjölk dracks* *Man drack mjölk*
Utelämning av *ha/har/att/som*	Se avsnitt 2.3.3
Satsflätor	Se avsnitt 3.6.7
Indirekt objekt – prepositionsobjekt	*Hon visade honom huset* *Hon visade huset för honom*
Koordinationsstrykning	*N sjunger och N spelar gitarr* *N sjunger och spelar gitarr*
Gapping (’verbellips’)	*Nisse åt korv och Pelle åt bröd* *Nisse åt korv och Pelle bröd*
Prepositionsstrandning (rektionsframflyttning)	*De lyssnar på musik* *Musik lyssnar de på [-]*
Lika NP-strykning	*Olle lovade att han skulle komma* *Olle lovade [-] komma*
Dubbelt fundament	*Kalle han kom sedan (jfr adjunktionellt så)*
Svansdubblering	*Kalle kom sedan, han.*

Tabell 47 Andra vanliga meningsbevarande transformationer finns bl.a. beskrivna i den forskning som följde den transformationsgrammatiska forskningen.

Denna kapiteldel har visat att analysen kan användas för omformulering av svensk text. Syntaktiska omflyttningar kan spegla en del av den frihet med vilken språkanvändningen rör sig och peka på mångfalden av former för samma innehåll. Detta kan ses som en negativ aspekt hos svenska, t.ex. från ett informa-

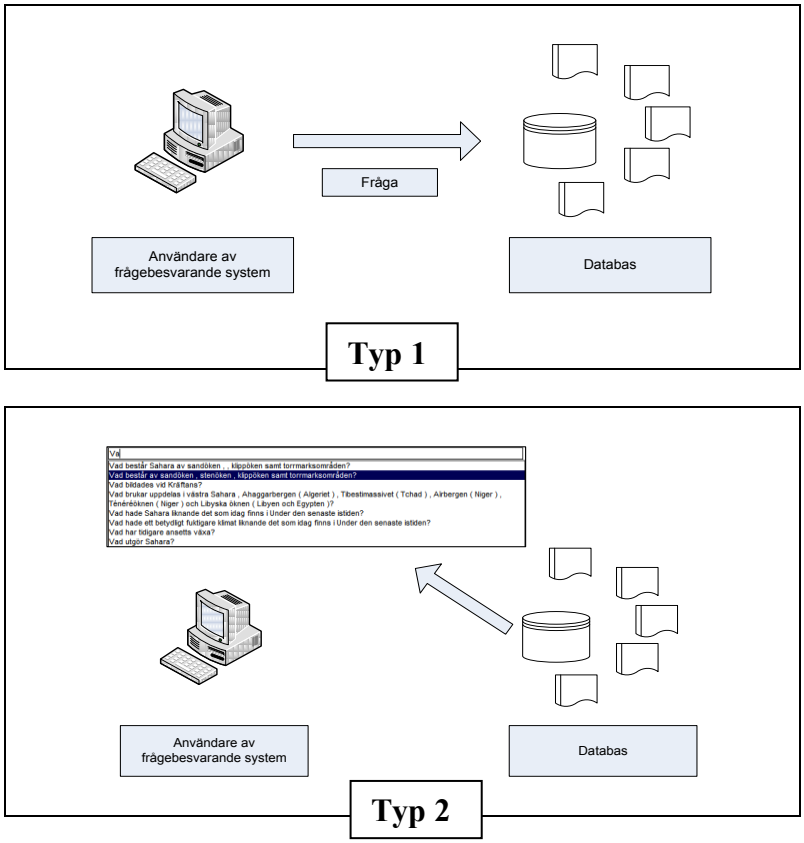
tionssökningsperspektiv. Det blir tydligt att ett meddelande kan formuleras just så skiftande även med samma ord. Kanske speciellt den syntaktiska variationen genom spetsställning visar hur svåra förutsättningarna är för *mönstermatchning* av t.ex. nexusrelationen subjekt – predikat i svenska, jämfört med t.ex. engelska. Det är med andra ord svårt att skriva en regel för flerordsmönstermatchning på svenska, eftersom det är oklart t.ex. var subjekt finns, jämfört med engelska med sin fixerade subjekt-finit-ordning. Att åstadkomma informationssökningsfunktionalitet för fri svensk text, fast med andra medel, är just syftet för den funktionalitet som beskrivs i nästa kapiteldel.

5.2 Automatisk generering av besvarade frågor från text

I ett frågebesvarande system med naturligt språk som gränssnitt (s.k. *natural language query systems*) tillåts en användare formulera frågor på naturligt språk som systemet med hjälp av en databas försöker besvara. En komplicerande aspekt för användaren av ett sådant system kan vara att denne i flera fall tvingas *gissa* huruvida en fråga kan besvaras av en viss databas. Dessutom måste ofta själva formuleringen vara matchande, med andra ord tolkas rätt och innehålla rätt ord, för att verkligen ge rätt svar i vissa befintliga system. Dessa svårigheter verkar speciellt hänga samman med en dold (*black box*) arkitektur som använder icke-trivial mönstermatchning, eller med system med oöverskådlig databas som t.ex. hela Internet. I denna kapiteldel beskrivs ett frågebesvarande system som med den beskrivna syntaxanalysen förändrar och eventuellt kan förenkla denna situation för användaren. Från ett användarperspektiv är tillämpningen anpassad så att systemet *explicit* genererar frågor som besvaras av en text när den läses in. Jämfört med ett system med en databas varemot en användare utan insikt om kunskapsbasen och/eller sättet som informationen är formulerad på ska ställa sina frågor, är idén här att i ett gränssnitt *visa* möjliga frågor – och inte tillåta andra frågor från användaren.

Sneiders (2002) delar upp system inom området *Natural Language Query Systems* i *AI-metoder* samt *mallbaserade metoder*. I dennes avhandling presenteras en mallbaserad metod vilket kan ha att göra med det faktum att engelska där hanteras och att engelska har en relativt fixerad ordföljd. Indelningen skulle möjligen innebära att det föreliggande systemet platsar i AI-kategorin, men ambitionen om artificiell intelligens känns inte relevant.[85]

[85] Även om den språkteknologiska applikation som beskrivs här skulle kunna klassificeras som artificiell intelligens är det i så fall en form av *svag AI* (*weak AI/narrow AI/applied AI*), dvs. en tillämpning som på sin höjd matchar mänsklig intelligens. Det är t.o.m. så att ansatsens medvetna relativa avsaknad av deduktion hittills i frågegenereringen, i strävan efter hög precision, motverkar förutsättningarna för så kallad *strong AI*.



Figur 44 I frågebesvarande system innebär ofta användningen att en fråga, som i Typ 1 ovan måste formuleras av användare – utan att denne känner till vilken information som databasen innehåller. I det föreliggande arbetet har systemtypen som kan illustreras som i Typ 2 utforskats. Denna variant innebär att de frågor som databasen besvarar genereras *explicit*. När en text öppnas med det beskrivna programmet skapas en lista med frågor som kan genomsökas av användaren. Detta med ett auto-kompletteringsgränssnitt över möjliga frågor. Arkitekturen behöver inte innebära en utspridd databas eller en *client-server*-lösning vilket bilden möjligen antyder.

Detta arbetes huvudfokus faller inom ramarna för syntaxområdet, men i och med identifikationen av funktionella syntaktiska led och de direkta tillämpningar som blir möjliga att genomföra därifrån, omformulering och frågegenerering, kan verka semantiska. I detta kapitel beskrivs funktionalitet för en tillämpning inom området *informationsextraktion*[86] på dessa grunder. Målsättningen är inte att fylla en databas med på förhand efterfrågade fakta utan att, med den beskrivna analysen som utgångspunkt, stringent samla information som finns på huvudsatsni-

[86] Denna tillämpning hör hemma inom *information extraction* (IE), vilket är ett delområde där relevant information från texter frambringas för en användare. Detta område måste skiljas från *information retrieval* (IR) som innebär sökning efter relevanta *dokument*, t.ex. på Internet. IE-system anses generellt vara mer komplicerade att bygga och inbegriper oftare en textanalysnivå som tar hänsyn till mer än bara ingående ord utan struktur (ord utan struktur: *'bag-of-words'*), vilket ofta är fallet inom information retrieval.

två. En korrekt funktionell huvudsatsanalys för en textmening bereder väg för att tekniskt använda propositionen (huvudsatsbetydelsen) för en sådan applikation. Målsättningen här är inte att lagra denna urvunna information i något speciellt mellanformat – eller att formulera den semantiskt otvetydigt. Istället kommer här att undersökas möjligheten att direkt från den grammatiskt funktionellt analyserade texten generera frågor som besvaras av den. Texten och systemet är alltså utformat för godtycklig text och information, utan att söka efter någon specifik uppgift i texten. Vad som samlas in är frågor som motsvarar *frågbara* primära satsled enligt analysen, här främst de obegränsade: subjekt, objekt/predikativ och adverbial. Att en texts informationsinnehåll skulle vara 'de frågor som den kan besvara' blir i detta fall relativt konkret. Detta tillvägagångssätt verkar inte ha någon direkt motsvarighet för svenska.

I Tabell 1 först i inledningskapitlet (sida 1) visades hur det var med hjälp av *frågor* som de olika leden kunde definieras i en grammatisk övning, men också att detta var tvunget att ske på form- och ordningsmässiga grunder för en datoranalys. Om det nu för en dator på detta sätt kan fastställas vilka grammatiska led som finns i en textmening, så kommer detta kapitel att ställa frågan om det är möjligt att gå åt andra hållet – dvs. att rent praktiskt generera de frågor som besvaras av en textmening eller av en hel text.

Struktur i kapiteldel 5.2, Automatisk generering av besvarade frågor från text

I avsnitt 5.2.1 beskrivs vilka typer av frågor som är tänkbara att generera utifrån en text med tanke på den aktuella analysnivån, och vilka möjliga frågor som ligger utanför ambitionen just nu.

I avsnitt 5.2.2 finns ett inledande exempel på hur frågegenerering kan ske via spetsställning av en funktionellt syntaktiskt analyserad text.

I avsnitt 5.2.3 tas upp frågan om hur frågor om olika satsled kan samlas in rent praktiskt. Det innebär också ett mindre försök till en mappningsfunktionalitet från satsled till motsvarande *hv*-fråga, men också andra syntaktiska sätt att omfråga textinformation.

Avsnitt 5.2.4 är en beskrivning av en faktisk implementation av en användarcentrerad funktionalitet som möjliggör frågegenerering när en text öppnas. Syftet är att snabbt leda användaren till tillgängliga formuleringar av en frågeställning om textinnehållet genom autokompletterande textinput. Den öppnade texten, som kan vara en artikel ur *Wikipedia*, finns samtidigt synlig för användaren och en frågeställning gör att textmeningen som innehåller svaret på frågan scrollas fram.

Avsnitt 5.2.5 innehåller resultatbeskrivning och ett försök till jämförelse med liknande system för svenska, även om klart jämförbara system verkar saknas. Dessutom redovisas tester som behandlar synonymutbyte för att på så sätt öka mängden frågeformuleringar.

5.2.1 En skiss av automatisk frågegenerering

Nedanstående korta text i Ex 70 är hämtad från artikeln *Kärnfysik* i svenska *Wikipedia*.[87]

Ex 70 a) Kärnfysik, den del av fysiken som berör atomkärnorna, deras beståndsdelar, struktur, dynamik och de krafter som verkar på och inom dem.
b) En atomkärna består av protoner och neutroner.
c) Dessa är i sin tur är uppbyggda av mindre partiklar, så kallade kvarkar.
d) Eftersom protonerna är positivt laddade och neutronerna saknar laddning, verkar elektromagnetiska krafter för att slita isär kärnan, som dock hålls ihop av stark växelverkan.

Dessa textmeningar representerar lite olika förutsättningar när det gäller att automatiskt generera frågor som ådagalägger den information de innehåller.

Ex 70 a) är speciell och typisk för inledningar av artiklar i encyklopedier genom att inte ens vara en huvudsats utan ha formen av en eller flera nominalfraser. De övriga meningarna består av huvudsatser.

Ex 70 b) kan tänkas besvara frågorna *Vad består en atomkärna av?* och *Vad består av protoner och neutroner?*

Ex 70 c) innehåller en anaforisk referens, *Dessa*. Genom en analog frågegenerering skulle bl.a. frågan *Vad är dessa (i sin tur) uppbyggda av?* tas fram. För att istället generera *Vad är protoner och neutroner (i sin tur) uppbyggda av?* krävs således *anaforlösning*.

Ex 70 d) innehåller (åtminstone) information för att besvara *Varför verkar elektromagnetiska krafter för att slita isär kärnan, (som dock hålls ihop av stark växelverkan)?, Vad verkar för att slita isär kärnan, (som dock hålls ihop av stark växelverkan)?* samt *För vad verkar elektromagnetiska krafter (eftersom protonerna är positivt laddade och neutronerna saknar laddning)?*

[87] Wikipedia (Svenska Wikipedia u.d.) – senast kontrollerad 2009-06-26. Ett skrivfel har rättats till här.

Utöver de nämnda frågorna kan även ja/nej-frågor (V1-frågor) enkelt genereras: *Består en atomkärna av protoner och neutroner?* osv. Så länge negerande adverbial (eller negerande NP-attribut såsom *ingen*) saknas blir svaren jakande, vilket gör dessa frågor lite mindre intressanta.

Ovanstående frågeexempel svarar genomgående mot informationsinnehåll som hänger samman med huvudsatsdelar och svaren motsvaras (förutom i fallet ja/nej-frågor) av en full primär satsdel, här subjekt, objekt och adverbial.

En mer raffinerad innehållsanalys skulle kunna innebära svar på frågor som (från Ex 70 c) *Vad kallas de mindre partiklar protoner och neutroner är uppbyggda av?*, och (från Ex 70 d) *Vad hålls kärnan (dock) ihop av (fast elektromagnetiska krafter verkar för att slita isär den)?*[88] Gemensamt för dessa exempel är att deras svar snarast motsvaras av underordnade attributiva led i den syntaktiska strukturen (*så kallad*-konstruktion respektive relativbisats), dvs. en djupare analys än den som huvudsakligen beskrivits i föregående kapitel.

5.2.2 Övergripande beskrivning av frågegenerering och spetsställningsparafraser

Frågegenerering kan göras enligt nedanstående övergripande procedurbeskrivning vilken har som syfte att skapa de nämnda frågorna (och i någon mening motsvarande svar) för uppgiften att möjliggöra att ställa frågor till en text. Det första som sker när en text läses in är att enkla huvudsatser extraheras från textmeningar. Varje samordnad, deklarativ huvudsats och samordnad primär finit verbfras ger potentiellt upphov till en enkel huvudsats.

- Uppdelning i huvudsatser för varje enkel huvudsats och varje primär finit verbfras, där generellt *subjekt* hämtas från föregående primära konjunkt som har subjekt. Dessa får därmed alla formen av en huvudsats. När en primär finit verbfras är samordnad är det nästan alltid subjektet som skiljer den från en huvudsats. Liknande konjunkter kan dock uppstå när andra ledtyper är vad som ärvs från den första konjunkten som i Ex 71, där den andra satskonjunkten åtminstone *kan* tolkas *I Schweiz var resultatutvecklingen positiv.*

  Ex 71 I Schweiz är marknadsläget stabilt och resultatutvecklingen var positiv. (hc04-147)

  1) För varje huvudsatsproposition: Generering av motsvarande ja/nej-fråga.

[88] Från den icke satsformade Ex 70 a) skulle naturligtvis frågan: *Vad är kärnfysik?* vara passande. Ett system för praktisk användning skulle givetvis kunna dra nytta av en så frekvent icke satsformad start i encyklopediska artiklar och skapa sådana frågor med ad hoc-metodik.

2) För varje huvudsatsproposition: Generering av alla parafraser genom spetsställning av varje spetsställbart primärt led

- För varje sådan konstellation (parafras) med spetsställt led: Generering (om möjligt) av motsvarande *hv*-fråga där det aktuella spetsställda ledet efterfrågas. Det betyder att spetsställning kan ske av led som inte borde kunna spetsställas enligt föregående kapiteldel. Begränsningarna här kan dock hindra att en sådan fråga genereras.

Ovanstående metodik behandlar en typisk textmening på följande sätt. Från textmeningen *Atomerna har ändå egenskaper som är likartade och kan emellertid undersökas samtidigt* genereras huvudsatspropositionerna i Ex 72 där den samordnade verbfrasen ärver subjekt från föregående konjunkt med subjekt.

Ex 72 a) Atomerna har ändå egenskaper som är likartade
b) Atomerna kan emellertid undersökas samtidigt

För att generera V1-formade ja/nej-frågor placeras fundamentledet på en annan acceptabel position (helst den kanoniska enligt satsschemat): *Har atomerna [ändå] egenskaper som är likartade?* samt *Kan atomerna [emellertid] undersökas samtidigt?* De frågor som genereras ska kunna besvaras positivt. När det gäller adverbialen *ändå* och *emellertid* är dessa här möjliga (eller rentav nödvändiga) att ta bort och fortfarande generera en fråga som besvaras positivt. Detta förhållande beror dock specifikt på de aktuella adverbialen och gäller t.ex. inte för satsadverbialet *inte*. I implementationen har ja/nej-frågorna lämnats därhän. Den följande beskrivningen gäller istället frågor som besvaras av primära led.

Spetsställning i de två huvudsatspropositionerna genererar följande uppsättningar av parafraser med bevarade sanningsvillkor.

- *Atomerna har ändå egenskaper som är likartade* ger upphov till följande konstellationer:

Ex 73 a) *<u>Atomerna</u> har ändå egenskaper som är likartade* (spetsställt subjekt)
b) *<u>Ändå</u> har atomerna egenskaper som är likartade* (spetsställt satsadverbial)
c) *<u>Egenskaper som är likartade</u> har atomerna ändå* (spetsställt objekt)

- *Atomerna kan emellertid undersökas samtidigt* ger upphov till följande konstellationer:

Ex 74 a) *<u>Atomerna</u> kan emellertid undersökas samtidigt* (spetsställt subjekt)
b) *<u>Emellertid</u> kan atomerna undersökas samtidigt* (spetsställt satsadverbial)
c) *<u>Undersökas</u> kan emellertid atomerna samtidigt* (spetsställt infinit verb)
d) *<u>Samtidigt</u> kan emellertid atomerna undersökas* (spetsställt tidsadverbial)

De ovanstående parafraserna motsvarar *hv*-frågor enligt följande uppställning utgående från de två huvudsatspropositionerna.

Parafras (med spetsställt led som aktuellt frågeämne)	Motsvarande *hv*-fråga
Atomerna har ändå egenskaper som är likartade (Subjekt)	*Vad/vilka har [ändå] egenskaper som är likartade?*
Ändå har atomerna egenskaper som är likartade (Satsadverbial)	*(Ändå* genererar ingen fråga)
Egenskaper som är likartade har atomerna ändå (Objekt)	*Vad har atomerna ändå?*

Atomerna kan emellertid undersökas samtidigt (Subjekt)	*Vad/vilka kan [emellertid] undersökas samtidigt?*
Emellertid kan atomerna undersökas samtidigt (Satsadverbial)	*(Emellertid* genererar ingen fråga)
Undersökas kan emellertid atomerna samtidigt (Infinit verb)	*Vad kan emellertid atomerna [göras] samtidigt? (*Ett *göras* kan klargöra vilket led som eftersöks)
Samtidigt kan emellertid atomerna undersökas (Tidsadverbial)	*När kan [emellertid] atomerna undersökas?*

Tabell 48 *Hv*-frågor utgående från en huvudsats med olika spetsställda primära led kan i många fall göras med en mappning, helt enkelt från ledets första ord till *hv*-ord. I uppställningen ingår, till skillnad från faktisk implementation, också fråga om det icke-finita verbet *undersökas*.

På det ovan beskrivna sättet ger meningen genom sina två huvudsatspropositioner upphov till minst fem *hv*-frågor (och två ja/nej-frågor). När det gäller *svar* på de frågor som genereras så modifieras inte den ursprungsmening som ger upphov till en fråga i den aktuella implementationen. Systemet genererar alltså inte en speciell svarsform, med elliptiskt uttryck, spetsställt led etc. Det finns en särskild poäng med att för användaren visa originaltexten och enbart grafiskt markera den mening som innehåller svaret, nämligen att i denna implementation överlåta vissa frågor som anaforlösning åt användaren (se nedan).

5.2.3 Satsled och motsvarande *hv*-frågetyper

För frågegenerering med avseende på de nämnda primära satsleden spetsställs de primära leden ett efter ett, enligt metoden ovan. Varje sådan uttryckt huvudsats kan, om det spetsställda ledet tillåter det, kopplas till en primärledsfråga (*hv*-fråga/frågeordsfråga). Uppgiften som här undersöks är hur väl ledets form (oftast huvudord, samt rektionens huvudord i PP) kan ge rätt fråga i en implementerad funktion.

Någon någorlunda heltäckande uppställning av vilka *hv*-frågeord som kan formas av olika led, inklusive prepositionsfraser eller av adverbial- och bisatser, på formmässiga grunder verkar hittills inte finnas för svenska i litteraturen. Nedanstående uppställningar återger implementerad metod för denna uppgift och har formen av ommappningsfunktioner.

5.2.3.1 Frågor om primära nominala led: subjekt och objekt/predikativ

Frågor som gäller fulla nominala led har generellt inledningen *vad/vem/vilken/vilket/vilka*.

Ex 75 a) Datorn är bara startraketen. → Vad är bara startraketen? (jb10-006)
b) Matti stank hel och hållen. → Vem stank hel och hållen? (kk60-065)
c) Landshövding, länsråd och biträdande länsråd är män. → Vilka är män? (hb14d-009)

Utöver dessa vanliga NP-strukturer uppbär dessutom bl.a. en delmängd av bisatserna regelmässigt roller som subjekt och objekt/predikativ, t.ex. *att*-bisatser och motsvaras företrädesvis av *Vad*.

En intressant aspekt hos frågor om nominala led är den potentiella ambiguitet rörande grammatiska funktioner de kan innebära, enligt Ex 76 och Tabell 49.

Ex 76 Bildningar av detta slag (SUBJ) förutsätter geografisk närhet (OBJ). (ja09-093)

Spetsställd version	Motsvarande fråga
(Original, samma form, dvs. ej spetsställd) *Bildningar av detta slag (SUBJ) förutsätter geografisk närhet (OBJ).*	*Vad (SUBJ) förutsätter geografisk närhet (OBJ)?*
Geografisk närhet (OBJ) förutsätter bildningar av detta slag (SUBJ).	*Vad (OBJ) förutsätter bildningar av detta slag (SUBJ)?*

Tabell 49 Spetsställningar och parafraser utgående från Ex 76.

Denna sats som ger upphov till två frågor kan nu jämföras med en annan sats (Ex 77) som innehåller samma ord och som formmässigt sammanfaller med parafrasen av den första.

Ex 77 Geografisk närhet (SUBJ) förutsätter bildningar av detta slag (OBJ).

Denna sats skulle ge upphov till samma spetsställningsversioner och uppsättning frågor, rent formmässigt sett, men rollerna subjekt och objekt är ombytta. Det finns med andra ord en flertydighet i vad (*vilken grammatisk funktion*) som efterfrågas med frågorna. Det är dock möjligt att åtminstone delvis entydigt efterfråga subjekt och objekt genom att skapa frågorna på underordnad nivå, i en relativsats, bryta ut det efterfrågade ledet och spetsställa det. Ex 78 och Ex 79 visar detta först utgående från Ex 76.

Ex 78
a) **Bildningar av detta slag (SUBJ) förutsätter geografisk närhet (OBJ).**
b) Det är bildningar av detta slag som förutsätter geografisk närhet *(Utbrytning av subjektet)*
c) Bildningar av detta slag är det som förutsätter geografisk närhet *(Spetsställning av det utbrutna subjektet)*
d) **Vad är det som förutsätter geografisk närhet?** *(Frågeomvandling från det spetsställda ledet)*

Motsvarande frågegenerering av *objektet*, också från Ex 76, kan ske på liknande sätt.

Ex 79
a) **Bildningar av detta slag (SUBJ) förutsätter geografisk närhet (OBJ).**
b) Geografisk närhet (OBJ) förutsätter bildningar av detta slag (SUB) *(Spetsställning av objektet)*
c) Det är geografisk närhet [som] bildningar av detta slag (SUB) *förutsätter (Utbrytning av objektet)*
d) Geografisk närhet är det [som] bildningar av detta slag (SUB) förutsätter *(Spetsställning av det utbrutna objektet)*
e) **Vad är det [som] bildningar av detta slag (SUB) förutsätter?** *(Frågeomvandling från det spetsställda ledet)*

Ex 78 och Ex 79 ovan visar hur det faktum att en underordnad sats ordföljd är mer fixerad leder till entydig tolkning i fråga om vilken grammatisk funktion frågan gäller. *Vad är det [som] bildningar av detta slag (SUB) förutsätter?* är entydig, och likaså *Vad är det som förutsätter geografisk närhet (OBJ)?*

Även om det är möjligt att konstruera frågor enligt ovan kan det också konstateras att en frågegenerering *utan* entydiggörande innebär att en eventuell felaktig syntaxanalys där subjekt och objekt förväxlats inte leder till något synbart fel. Det beror på att frågan i sig är flertydig, men denna eventuellt felaktiga analys tydliggörs ”i onödan” för en användare ifall utbrytning från relativsats används som i Ex 78 och Ex 79.

5.2.3.2 Frågor om adverbiella led

Frågor gällande adverbialled bjuder på lite andra förutsättningar. Adverbialfrågor där adverbialet är en PP kan först och främst ställas på två vanliga sätt, som *hv-frågor* (*Hur går båten?*) eller *pied piping-version* (*Med vad går båten?*) [89] där båda kan motsvara adverbialdelen av *Båten går med ångkraft*. Här råder förhållandet att pied piping-versionen är enklare att generera (samma preposition plus ett nominalt *hv*-ord som motsvarar komplementsdelen), medan frågetypen med *hv*-ord, å andra sidan, kräver mindre kännedom om systemets kunskapsbas från en användares sida: Att fråga *hur* kräver inte kunskap om att svaret inleds med just *med*. Det antas vara enklare för användaren att finna en befintlig frågegenerering på så sätt.

Med-PP i denna instrumentella tolkning är vanlig och *med*-PP motsvaras ofta av *hur*-frågor. Andra prepositionsfrastyper, t.ex. de som inleds med *i* eller *på,* kan motsvaras av betydligt fler frågetyper. Adverbial i form av bisatser har dock ett mer entydigt förhållande till motsvarande frågetyper. Det typiska förhållandet mellan spetsställbarhet och frågbarhet verkar vara att omfrågbara led kan spetsställas men att det omvända inte gäller i samma utsträckning.

- Huvuddelen av de olika förekomsterna av adverbialled är möjliga att både spetsställa och fråga om.

- En grupp adverbial är möjliga att spetsställa men är inte omfrågbara på ett enkelt sätt. Det gäller typiskt satsadverbial: *inte, trots allt, ändå.*

- En i sammanhanget liten grupp utgörs av de adverbial som varken kan omfrågas eller spetsställas. Lindberg och Svensson (1992) pekar ut *ju, väl, verkligen, också, ej.*

5.2.3.3 *Hv*-fråga, rektionsfråga eller *pied piping*

Led som är PP-formade adverbial kan omfrågas genom tre grundläggande frågestrukturer. Frågan som gäller detta led, t.ex. *till närmsta tjänsteman* i satsen *Ärendet skulle skickas till närmsta tjänsteman* kan motsvaras av 1) *hv*-ord (*Vart skulle ärendet skickas*), 2) rektionsframflyttning med prepositionsstrandning (*Vem skulle ärendet skickas till* [_]) eller 3) s.k. pied piping (*Till vem skulle ärendet skickas* [_]).

[89] 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).

I detta sammanhang blir valensinformation för verb användbart: De här tidigare omdiskuterade prepositionsobjekten erbjuder nämligen jämförelsevis begränsade möjligheter till omfrågning – dessa led verkar ofta inte kunna representeras av ett enskilt *hv*-ord, enligt Tabell 50. Denna begränsning i frågbarhet hos de flesta prepositionsobjekten verkar nästan gälla som en definierande aspekt.[90]

Version med spetsställd PP	*HV*-ord	Rektionsframflyttning med strandad prep	Pied piping
Till närmsta tjänsteman skulle ärendet skickas	*Vart skulle ärendet skickas?*	*Vem skulle ärendet skickas till?*	*Till vem skulle ärendet skickas?*
Efter matchen gick de	*När gick de?*	*Vad gick de efter?*	*Efter vad gick de?*
På musik lyssnade de alltid	-	*Vad lyssnade de alltid på?*	*På vad lyssnade de alltid?*
Åt de fattiga ger Röda Korset	-	*Vem ger Röda Korset åt?*	*Åt vem ger Röda Korset?*

Tabell 50 De adverbial som är fyllnadsled (prepositionsobjekt), dvs. de i de två sista fallen, verkar generellt sakna möjligheten till helt motsvarande *hv*-fråga.

Agentadverbial verkar ha samma begränsning som prepositionsobjekten ofta har på svenska, dvs. de saknar enkel motsvarighet av något enskilt *hv*-ord (? *vemav sågs tomten?)*. I aktuell version särbehandlas dock ej dessa adverbialtyper.

5.2.3.4 Ledmappning till frågeord

Som nämnts ovan verkar det inte finnas någon tidigare någorlunda fullständig uppställning över förhållandet mellan inledningsord i bisatser eller prepositionsfraser till motsvarande frågeord i svenska i litteraturen. Den uppställning som här presenteras är en medvetet tentativ sådan grundad på samtliga ord som förekommer som bisatsinledare och prepositioner vid minst ett tillfälle i SUC 2.0. Härutöver kommer de *hv*-ord som svarar mot de nominala leden (*vad, vem, vilket* etc.). De olika grupperna av ord med specificerade ordklasstaggar innehåller

[90] Denna poäng har i alla fall inte påträffats i läst grammatiklitteratur. Några teoretiska exempel (delvis med pseudo-ord) utgående från faktisk valensinformation från NEO: *råskälla - på ngn: ?Vempå råskällde du?, skoja: med ngt: ?Vadmed skojade du?, Strunta: i ngt: ?Vadi struntade du?* Däremot verkar följande möjligtvis fungera: *upphöja:& ngn/ngt till ngn/ngt: Vadtill upphöjde du honom?* Frågan är om bristen på *hv*-ordmotsvarighet handlar om avsaknad av passande ord eller om kanske avsaknaden av ord från början kommer sig av att de två andra formerna är så dominerande för frågor om prepositionsobjekt.

både ord som motsvarar nominala led och andra ord som när de inleder motsvarar adverbiella led (exempelvis hos subjunktioner: *att/eftersom*). I många fall av vanliga PP är det dock inte enbart inledningsordet som avgör fråga utan en kombination av inledningsord och rektionens huvudord (*på bordet/på medeltiden*), vilket är ett allmänt försvårande faktum.

Det är viktigt att klargöra att uppställningen gäller led där det aktuella ordet inleder: många av bisatsinledarna fungerar även som relativsatsinledare, dvs. de inleder där bara efterställda attribut, men dessa fall är ju inte aktuella om syntaxanalysen är korrekt. *Frågande/relativa adverbial* (*HA* i SUCs kodning) blir ett tydligt exempel på hur en satsledsinledning med ett sådant ord inte är just adverbial utan uppbär nominal funktion (*Hon undrade var de var → Hon undrade vad?*). Uppställningen här är gjord med avseende på ordklasstagg (enligt SUC 2.0) hos det inledande ordet. Bestämningen av de motsvarande *hv*-orden är delvis grundad på enheter med faktiska förekomster med hjälp av en speciell sökfunktionalitet i programmet som kan lista de s-enheter som innehåller en viss ordform.

Det har nämnts att analysformatet för parsningen i detta arbete är mindre detaljerat än t.ex. i Mamban eftersom adverbialtyper ej specificeras. Om mappningen till *hv*-ord skulle ske felfritt skulle den dock innebära en ännu mer detaljerad adverbialbeskrivning (*rumsadverbial* svarar t.ex. mot *var, vart* etc.).

Bisatsformade adverbiella och nominala led

Kategorin explicita bisats- och relativsatsinledare svarar mot en handfull ordklasstaggar enligt SUC 2.0 och inleder både nominala och adverbiella led. De olika ordklasstaggarna är *SN, HS, HA, HD, HP* med olika särdrag. Subjunktioner (SN) är de typiska bisatserna medan de andra också fungerar som inledning på efterställda attribut, men dessa utgör ju inte egna led och hanteras därför ej. De led som inleds med en H-ordklass (*HS, HA, HD, HP*) får genomgående en motsvarighet i en *vad*-fråga: *Var (HA) de var verkade oklart*: *Vad verkade oklart*? Uppställningen över subjunktioner (*SN*) och motsvarande *hv*-frågeord som använts i dessa pilottester återfinns i Appendix.

PP-formade adverbial

Det förekommer 133 olika ord i SUC 2.0 som på minst ett ställe är taggade som preposition. Uppställningen gäller dessa ord som i många fall är stavningsvarianter (*över/öfwer*). Antalet är i rimlig överensstämmelse med Allén (1972) som räknar med 200–300 olika prepositioner i svenska. Sjöström (1985) förklarar

detta stora antal prepositioner i svenska med möjligheten till prepostionsformationer som är sammansatta av adverb och preposition (*uppemot, frampå*).

För PP-adverbial gäller alltså möjligheten att de kan vara prepositionsobjekt och då ofta saknar motsvarighet i enskilt *hv*-frågeord. De svårigheter som finns för korrekt frågegenerering är dock i hög grad snarare beroende på att listning av *huvudorden* (rektionens huvudord i PP) med olika typer av *hv*-ordmotsvarighet är ofullständig. Den listning av prepositioner till *hv*-ord som görs i detta arbete har använts för att ge en bild i den följande testutvärderingen av prototypen. Mappningen återfinns i Appendix.

Adverbfrasformade och NP-formade adverbial

Huvudord som är adverb har ej listats men denna grupp antas kunna täckas i hög grad givet mer tid. Många satsadverb svarar emellertid inte enkelt mot något *hv*-ord. När det gäller NP-formade adverbial ges dessa en heuristisk frågeordsmappning till *när*. Det antas vara relativt enkelt att specificera även dessas motsvarande frågeord bättre givet mer tid.

5.2.4 En implementation av frågegenerering mot *Wikipedia* eller valfri text

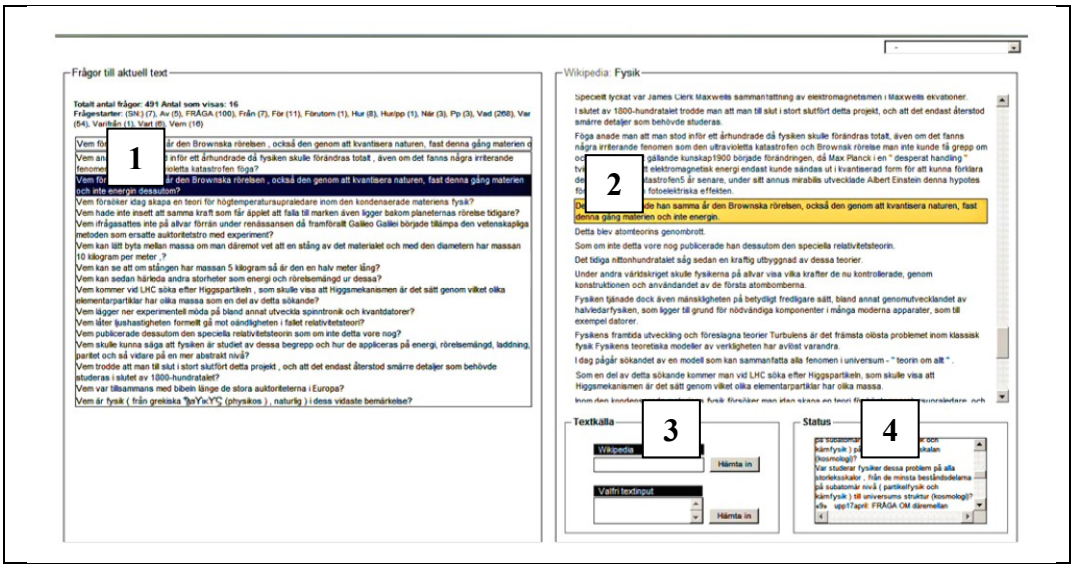
Det följande avsnittet behandlar hur ett system för automatisk frågegenerering enligt ovan ter sig i praktiken. Implementationen som gjorts använder sig av de föregående stegen i detta arbete och genererar frågor för en vald text.

Detta är den applikation som i projektet har flest beroenden. Korrektheten för ordklasstaggningen begränsar korrektheten för den heuristiska primära satslösningen, vilken i sin tur begränsar möjligheterna för korrekt spetsställning. Frågegenerering har en korrekthet som hänger på korrektheten hos alla föregående processteg. Det ska dock noteras att det är möjligt att förhöjd korrekthet hos efterföljande steg i denna kedja är möjlig. Framförallt är, som nämnts, schemaparsningens chunkning okänslig för många fel i särdragen i ordklasstaggar.

En särskilt användbar fri resurs med faktainnehåll som ansetts lämplig är den svenska *Wikipedia*, vars databas finns fri för nedladdning.[91] Användningsscena-

[91] Den flerspråkiga Internetbaserade encyklopedin *Wikipedia* är baserad på wiki-teknologi där användare kan skapa och redigera sidorna. Verket finns i över 200 språkversioner. Den engelska encyklopedin har i skrivande stund fler än två miljoner artiklar medan den svenska versionen har ca 332 000. Denna dynamiska encyklopedi har kritiserats på olika grunder, se t.ex. Spetz (2008), men kritik om innehåll och korrekthet är mindre relevant för det beskrivna syf-

riot för en mer utbyggd funktionalitet skulle kunna vara att användaren öppnar ett dokument eller en sida på en webbplats och vid öppnandet samtidigt får möjlighet att generera frågor till texten. Själva texten finns synlig hela tiden för användaren.



Figur 45 Gränssnittet i programmet upptas huvudsakligen av formulär för frågeval och själva texten.

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 Wikipedia eller annan textinput
4) Statusruta för diverse information under körning

Användningen av programmet, så som det ser ut i aktuell implementation, sker genom att en text matas in eller genom att en Wikipedia-artikel öppnas. Programmet lagrar Wikipedias textartiklar filvis lokalt. Vid öppnandet av en text/artikel frågas användaren om frågor ska genereras. Om frågor genereras visas artikeln tillsammans med en textruta med autokompletteringsfunktion där de tillgängliga frågorna finns. När en fråga väljs blir svarsmeningen i texten markerad och scrollas fram så att den blir synlig. Närmare bestämt så visas svarsmeningen med några ovanstående rader text för att på det sättet möjliggöra att användaren själv löser potentiell anaforisk referens och får annan kontext, se Figur 45. Programmet genererar enbart frågor för s-enheter som får en full analys, vilket alla inte får.

---

tet, medan det faktum att texter generellt är relativt grammatiskt korrekta är positivt, och att de är producerade av så många författare är intressant. Återkommande internationella konferenser som *Wikimania* och *WikiSym* uppmärksammar Wikipedias snabbt vunna betydelse som fri resurs.

5.2.5 Test av frågegenerering

För att testa hur väl frågegenerering fungerade i fråga om korrekta frågor (form på frågan, rätt frågeord osv.) med denna pilotversion av frågeordsmappning gjordes testningar mot slumpvalda artiklar från *Wikipedia*. Eftersom själva programmet saknar möjligheten att slumpvis välja artiklar från datorn lokalt så användes *Wikipedias* egen hemsida med funktionen 'slumpartikel'. Artikeltexten kopierades utan att ta med rubrik, tabeller osv. Eftersom programmets egen textdatabas med artiklar för närvarande inte skiljer ut dessa andra element lika bra, var utgångsläget mer gynnsamt under dessa försök. Här visas text och faktiska frågor från fyra framslumpade artiklar utan någon förskönings. Det behöver knappt påpekas att felen i hög utsträckning handlar om fel val av *hv*-ord, t.ex. *var* eller *när* för en *i*-PP. Ett liknande fel som borde kunna åtgärdas enkelt är att personer kallas *vad*. Prepositionsobjekt identifieras eller särbehandlas ej i testet. Det antas, i likhet med så gott som alla delmoment i detta avhandlingsprojekt, finnas en stor möjlighet till förbättring av dessa steg givet mer tid.

Det följande är ett mindre test med försiktigt satta inställningar mot fyra slumpvalda artiklar. I testet som här beskrivs räknas inte typografiska fel som skiftläge hos begynnelsebokstav eller införda mellanslag. Inte heller ses ett ospecificerat (dvs. ej anaforupplöst) *han* etc. som fel. De manuella markeringarna för genererade frågor i nedanstående undersökning av korrektheten i tabellerna ska tolkas på följande sätt (de frågor som är korrekta eller acceptabla markeras ej):

*	Fel på frågan givet den information som svaret ger, t.ex. fel frågeord eller syntaxfel,
?	Ej idealisk fråga (inte helt rätt) givet dess svar
(?)	Inte helt specificierad fråga (t.ex. *var/när*) som bara delvis är korrekt, givet svarets information. Detta beror på att fullständig listning av rektioners huvudord i PP saknas i denna prototyp.
(k)	Korrekt fråga med avseende på svaret, men på oavsedda grunder – svaret finns t.ex. i en annan del av textmeningen.

Kirk Franklin
Kirk Franklin, född 26 januari 1970 i Fort Worth, Texas, USA, är en amerikansk gospelmusiker och sångare. Franklin var uppväxt i kyrkan och tog tidigt till sig körmusiken han upplevde där. Han började spela piano när han var fyra år och han ledde sin första kör när han var elva. I tonåren gjorde han revolt och bröt upp från kyrkan. Han levde ett stormigt liv och umgicks i kriminella gäng. Efter att en vän blivit skjuten återvände han till Gud och kyrkan. Under tidigt 90-tal startade han upp en gospelkör som han kallade "The Family". Franklin har även samarbetat med andra körer genom åren som t.ex. God's Property. Kirk Fanklin har besökt Sverige två gånger. Den 10:e september 2008 spelade han på Conventum Arena i Örebro och den 8:e november 2009 i Globen i Stockholm.

Genererade frågor
(?) Från vad / från när bröt han upp?
(?) När (var) återvände han till Gud och kyrkan?
Vad började han spela när han var fyra år?
Vad gjorde han i tonåren?
* Vad har kirk fanklin besökt Sverige?* (*Fyra gånger* är adverbial)
Vad ledde han när han var elva?
? Vad levde han? (*Han levde ett stormigt liv*)
* Vad spelade han på Conventum i Globen i Stockholm den 10 :e september 2008?
Vad startade han upp under tidigt 90-tal?
Vad tog franklin tidigt till sig?
Vad var franklin i kyrkan?
Vad är kirk franklin , född 26 januari 1970 i fort worth , texas , usa ,?
* Var gjorde han revolt? (*I tonåren: När*)
* Var spelade han Arena i Örebro och den 8:e november 2009 i Globen i Stockholm den 10 :e september 2008? (Fel syntaxanalys: *på Conventum Arena* är ett sammanhängande adverbial)
* Var spelade han på Conventum Arena i Örebro och den 8:e november 2009 den 10 :e september 2008?
* Var startade han upp en gospelkör som han kallade The Family? *(Under tidigt 90-tal: När)*
Var umgicks han?
Var var franklin uppväxt?
Vart återvände han efter att en vän blivit skjuten?
Vem bröt upp från kyrkan?
Vem började spela piano när han var fyra år?
Vem gjorde revolt i tonåren?
Vem har besökt Sverige två gånger?
* Vem har kirk fanklin besökt två gånger? (*Sverige: Vad*)
Vem ledde sin första kör när han var elva?
Vem levde ett stormigt liv?
Vem spelade på Conventum Arena i Örebro och den 8:e november 2009 i Globen i Stockholm den 10 :e september 2008?
Vem startade upp en gospelkör som han kallade The Family under tidigt 90-tal?
Vem tog tidigt till sig körmusiken han upplevde där?
Vem umgicks i kriminella gäng?
Vem var uppväxt i kyrkan?
Vem är en amerikansk gospelmusiker och sångare?
Vem återvände till Gud och kyrkan efter att en vän blivit skjuten?

Tabell 51 Text och genererade frågor till Kirk Franklin från Wikipedia (Svenska Wikipedia u.d.).

Slaget vid Dak To
Slaget vid Dak To var ett slag mellan den amerikanska armen och den nordvietnamesiska armen (NVA) under Vietnamkriget som utkämpades vid den sydvietnamesiska-nordvietnamesiska gränsen från den 3 till 22 november 1967. Slaget blev ett av de blodigaste under hela kriget. Slaget började efter attacker av NVA på de amerikanska och sydvietnamesiska ställningarna längs gränsen, blodiga och intensiva strider rasade under tre veckors tid innan amerikanerna efter massiva förstärkningar och med flyg och artilleriunderstöd lyckas trycka tillbaka NVA. Även om NVA drevs tillbaka så blev slaget kostsamt för amerikanarna som förlorade nära 300 man döda och nästan 1000 sårade.

Genererade frågor
(?) När (var) började slaget på de amerikanska och sydvietnamesiska ställningarna längs gränsen?
Vad blev ett av de blodigaste under hela kriget?
* Vad blev slaget under hela kriget? (*Under hela kriget* är ej fråstående adverbial)
Vad började efter attacker av NVA på de amerikanska och sydvietnamesiska ställningarna längs gränsen?
* Vad började slaget efter attacker av NVA på de amerikanska och sydvietnamesiska ställningarna?
* Vad drevs tillbaka så blev kostsamt för amerikanarna som förlorade nära 300 man döda och nästan 1000 sårade även om nva? (Fel syntaxanalys)
Vad var ett slag mellan den amerikanska armen och den nordvietnamesiska armen (NVA) under Vietnamkriget som utkämpades vid den sydvietnamesiska-nordvietnamesiska gränsen från den 3 till 22 november 1967?
Vad var slaget vid dak to mellan den amerikanska armen och den nordvietnamesiska armen (NVA) under Vietnamkriget som utkämpades vid den sydvietnamesiska-nordvietnamesiska gränsen från den 3 till 22 november 1967?
(?) Var blev slaget ett av de blodigaste? (Svar: *under hela kriget*)
* Var började slaget efter attacker av NVA längs gränsen? (*på de amerikanska och sydvietnamesiska ställningarna är attribut till attacker*, ej adverbialled)
(k) Var rasade blodiga och intensiva strider?
(k) Var var slaget vid dak to ett slag mellan den amerikanska armen och den nordvietnamesiska armen (NVA)?
* Var var slaget vid dak to ett slag under Vietnamkriget som utkämpades vid den sydvietnamesiska-nordvietnamesiska gränsen från den 3 till 22 november 1967? (*Mellan…* är svarar ej mot *var*)
? Vilka rasade under tre veckors tid innan amerikanerna efter massiva förstärkningar och med flyg och artilleriunderstöd lyckas trycka tillbaka NVA? (*blodiga och intensiva strider: Vad*)

Tabell 52 Text och genererade frågor till *Slaget vid Dak To* från Wikipedia (Svenska Wikipedia u.d.).

Lauri Karvonen
Lauri Antero Karvonen, född 21 november 1952 i Uleåborg, Finland, finländsk statsvetare. Karvonen avlade politices doktorsexamen vid Åbo Akademi 1981 på en avhandling Med vårt västra grannland som förebild: en undersökning av policydiffusion från Sverige till Finland. Hans handledare var professor Dag Anckar. Efter en forskartjänst vid Århus universitet i Danmark kom han att verka vid Åbo Akademi och Finlands Akademi innan han utnämndes til professor i jämförande politik vid Bergens universitet 1994. Fyra år senare, 1998, var han tillbaka i Åbo som professor i statsvetenskap. Han har fungerat som handledare till flera namnkunniga finländska och svenska statsvetare, exempelvis Carsten Anckar. Hans forskning har främst rört sig inom jämförande politik och han har även gett ut flera läroböcker kring demokrati och statsskick. Han har även varit gästprofessor vid det svenska Demokratiinstitutet kopplat till Mittuniversitetet.

Genererade frågor
När kom han att verka vid Åbo Akademi och Finlands Akademi efter en forskartjänst vid århus universitet i danmark?
(?) Från vad / från när avlade karvonen politices doktorsexamen vid Åbo Akademi 1981 på en avhandling Med vårt västra grannland som förebild : en undersökning av policydiffusion till Finland?
* Hur avlade karvonen politices doktorsexamen vid Åbo Akademi 1981 på en avhandling som förebild : en undersökning av policydiffusion från Sverige till Finland? (*Med vårt västra grannland*: Titel)
(?) När (var) kom han att verka vid Åbo Akademi och Finlands Akademi innan han utnämndes til

professor i jämförande politik vid Bergens universitet 1994?
Vad avlade karvonen vid Åbo Akademi 1981 på en avhandling Med vårt västra grannland som förebild : en undersökning av policydiffusion från Sverige till Finland?
* Vad avlade politices doktorsexamen vid Åbo Akademi 1981 på en avhandling Med vårt västra grannland som förebild : en undersökning av policydiffusion från Sverige till Finland? (*Vem*)
Vad har han även gett kring demokrati och statsskick?
Vad har han även varit vid det svenska Demokratiinstitutet kopplat till Mittuniversitetet?
Var avlade karvonen politices doktorsexamen 1981 på en avhandling Med vårt västra grannland som förebild : en undersökning av policydiffusion från Sverige till Finland?
(k)Var avlade karvonen politices doktorsexamen vid Åbo Akademi 1981 Med vårt västra grannland som förebild : en undersökning av policydiffusion från Sverige till Finland? (*På en avhandling*)
* Var har han även gett ut flera läroböcker? (*kring demokrati och statsskick*, ej *var*)
Var har han även varit gästprofessor kopplat till Mittuniversitetet? (*vid det svenska Demokratiinstitutet kopplat till Mittuniversitetet* är ett enda adverbial)
Var har hans forskning främst rört sig?
Var kom han att verka innan han utnämndes til professor i jämförande politik vid Bergens universitet 1994 efter en forskartjänst vid århus universitet i danmark?
Var var han tillbaka som professor i statsvetenskap fyra år senare , 1998 ,?
* Vart avlade karvonen politices doktorsexamen vid Åbo Akademi 1981 på en avhandling Med vårt västra grannland som förebild : en undersökning av policydiffusion från Sverige? (från Sverige till Finland, del av titel)
* Vart har han fungerat som handledare? *(till flera: Till vem)*
* Vart har han även varit gästprofessor vid det svenska Demokratiinstitutet kopplat?
Vem har fungerat som handledare till flera namnkunniga finländska och svenska statsvetare , exempelvis Carsten Anckar?
Vem har även gett ut flera läroböcker kring demokrati och statsskick?
Vem har även varit gästprofessor vid det svenska Demokratiinstitutet kopplat till Mittuniversitetet?
Vem kom att verka vid Åbo Akademi och Finlands Akademi innan han utnämndes til professor i jämförande politik vid Bergens universitet 1994 efter en forskartjänst vid århus universitet i danmark?
(k)Vem var hans handledare professor Dag? (*Dag Anckar*: egennamn)
Vem var tillbaka i Åbo som professor i statsvetenskap fyra år senare , 1998 ,?
* Vilka har främst rört sig inom jämförande politik? (*Dag Anckar*: egennamn)
* Vilka var professor Dag Anckar? (*Dag Anckar*: egennamn)

Tabell 53 Text och genererade frågor till *Lauri Karvonen* från Wikipedia (Svenska Wikipedia u.d.).

Duane Eddy
Duane Eddy, född 26 april 1938 i Corning, New York, är en amerikansk rockmusiker och gitarrist. Eddy är den mest framgångsrike instrumentalisten i rockhistorien och har sålt mer än 100 miljoner skivor. Duane Eddy införde "twang" i rock'n'rollen. Hans låt "Rebel Rouser" från juli 1958 är en av 1950-talets mest populära instrumentala rock'n'roll-hits. Ett par populära uppföljare var "Peter Gunn" och "Shazam", båda från 1960. En nyinspelning av "Peter Gunn" med Art of Noise gav honom 1986 en Grammy för bästa instrumentala rock. Han valdes in i Rock and Roll Hall of Fame 1994.[1]

Genererade frågor
(?) Från vad / från när var ett par populära uppföljare Peter Gunn och Shazam , båda?
För vad gav en nyinspelning av peter gunn med art of noise honom 1986 en Grammy?
Vad gav en nyinspelning av peter gunn med art of noise 1986 en Grammy för bästa instrumentala rock?

Vad gav en nyinspelning av peter gunn med art of noise honom 1986 för bästa instrumentala rock?
* Vad gav en nyinspelning av peter gunn med art of noise honom en Grammy för bästa instrumentala rock?
Vad gav honom 1986 en Grammy för bästa instrumentala rock?
Vad införde duane eddy?
* Vad införde twang i rock'n'rollen?
* Vad valdes han in i Rock and Hall of Fame 1994 .[1]? (*Rock and Roll Hall of Fame*: en NP)
* Vad valdes han in i Rock and Roll 1994 .[1]? (*Rock and Roll Hall of Fame*: en NP)
Vad var Peter Gunn och Shazam , båda från 1960?
Vad var ett par populära uppföljare från 1960?
(?) Vad är duane eddy , född 26 april 1938 i corning , new york ,?
Vad är eddy i rockhistorien?
* Vad är en amerikansk rockmusiker och gitarrist?
Vad är hans låt rebel rouser från juli 1958?
* Var valdes han in Roll Hall of Fame 1994 .[1]? (*Rock and Roll Hall of Fame*: en NP)
Var är eddy den mest framgångsrike instrumentalisten?
Vem har sålt mer än 100 miljoner skivor?
Vem valdes in i Rock and Roll Hall of Fame 1994 .[1]?
Vem är den mest framgångsrike instrumentalisten i rockhistorien?
* Vilka är en av 1950-talets mest populära instrumentala rock'n'roll-hits? (*Vad*)

Tabell 54 Text och genererade frågor till *Duane Eddy* från Wikipedia (Svenska Wikipedia u.d.).

Resultatet visar att minst 54 av 95 frågor (ca 57 %) här antogs vara helt rätta och ytterligare 4 (ca 4 %) antas vara acceptabla i denna tidiga pilotversion. Tillsammans med de som räknas som dåligt specificerade (t.ex. *från vad/när*), vilka antagligen skulle kunna förbättras, utgör de här ca 2/3 av frågorna. Resultatet av denna mindre undersökning, visade att oväntat få felmarkeringar berodde på grava syntaxanalysfel. Felen rörde bl.a. om vilket frågeord som valts och eftersom den delen av projektet har ägnats relativt lite tid ger detta stora förhoppningar på förbättrad korrekthet givet mer tid.

Denna metod har varit tänkt att gagna *precision*-värdet i uppgiften. Med andra ord ligger fokus på att ge ett korrekt svar givet varje fråga. Naturligt är då att vad som kallas *recall*-värdet istället blir lägre: dvs. frambringandet av de frågor som texten kan anses besvara, eller till och med alla de formuleringar av frågor som är giltiga. Recall-värdet tillåts att bli högt i ett system som tillåter användaren att ställa vilken fråga som helst.

Att frågorna i viss mån inte är grammatiska kan också i vissa sammanhang ses som ett relativt lindrigt fel. Att en fråga som blir tillgänglig inte är korrekt uttryckt innebär inte desto mindre en möjlighet att använda den för att nå relevant information i texten, som ett sökverktyg. Programmet ger ju inte något falskt svar, eftersom användaren visas den del av texten som givit upphov till frågan. Ett vanskligare förlopp hade varit ett där ett svar automatiskt skulle genereras.

Om korrektheten för svaren sjunker är det troligen ett klart värre slag mot användarens förtroende för systemet. Systemet kan i en mening sägas ha 'sökegenskaper': det fungerar genom att söka fram rätt textmening – skillnaden mot vanlig sökning är just att söktermen är formulerad som en fråga. Genom att systemet bara söker och levererar en textmening givet frågan är det till syvende och sist upp till användaren själv att bedöma om det är en textmening som besvarar frågan. Eftersom ovanstående resultat kommer från en väldigt kort periods arbete antas metoden kunna användas med bra resultat i en längre tidsram för arbetet.

Möjliga säkra utökningar av mängden frågor givet aktuell analys

Om varje frågbart led skulle generera en unik fråga så skulle antalet frågor här vara lika med dessa antal led. I praktiken är så inte fallet, utan syntaxanalyser som leder till felaktiga analyser (t.ex. där huvudsatser missas) tas bort. Detta kan ses vara i linje med den i övrigt "försiktiga" ansatsen där *precision*-värde värderas högre än *recall*-värde. Det finns några tänkbara sätt att utöka mängden besvarade frågor eller över huvud taget formuleringar av frågor för att förbättra användbarheten. Ett första sätt vore att ta till vara propositioner som uttrycks på underordnade satsnivåer. Det betyder att förutsättningen är en djupare analys än den som förekommer hittills. Här gäller den viktiga distinktionen om huruvida verbet i matrissatsen är *faktivt* eller inte – dvs. skillnaden mellan det säkerställda sanningsvärdet för *tåget hade kommit* i *De visste att tåget hade kommit* respektive *De gissade att tåget hade kommit*. Om verbet är *faktivt* kan alltså *Tåget hade kommit* behandlas som en ytterligare huvudsats att generera frågor utifrån.

Ett andra sätt att utöka mängden frågor som verkar relativt säkert vore att producera 'attributfrågor'. Från en sats av typen *Han köpte den nyaste utgåvan*, kan själva attributet generellt sett efterfrågas: *Vilken utgåva köpte han?* Denna typ av analys verkar görbar med aktuell analysnivå.

Ett sätt att ge fler frågeformuleringar, som inte så tydligt är informativt annorlunda jämfört med de befintliga, handlar om versioner med borttagande av en grupp adverbial, typiskt satsadverbial, som på olika sätt handlar om talarkommentarer. *Emellertid* är ett typexempel som skulle tas bort *i Vad har Landstinget emellertid köpt?*.

När det gäller förändring av frågor genom utbyte av ingående termer till synonymer kan detta exemplifieras med att *De köpte en jycke* skulle kunna ge *De*

köpte en hund.[92] Om ordparet innebär en stark synonymirelation så är antagandet att ord kan bytas ut på vilken position som helst i en textmening. För denna typ av termabstrahering undersöktes svenska versionen av *WordNet* (*Svenskt OrdNät, SWordNet*) (Viberg, Lindmark, o.a. 2002).[93] I dessa försök används ingen mappningsfunktionalitet till grundformer, utan bara ordförekomster som finns i grundform i texten har kommit i fråga för utbyte.

På samma gång testades även *Folkets synonymordlista* (Kann och Rosell 2005), som skapats genom användarbedömningar av synonymer. Testet visar att rent synonymutbyte kräver relativt stark synonymi, vilket gör att de flesta synonymiförslagen (med lägre synonymipoäng på den omnämnda femgradiga skalan) i ordlistan blir oanvändbara just här, om det som här anses viktigt att inte introducera obesvarade frågor. Det är möjligt att expansion av frågemängden genom lexikala ordutbyten skulle kunna ske i begränsad omfattning, men det krävs vidare undersökningar för att ställa in begränsningar i programmet.

Undersökningen här visar vad som händer om förslag från de två källorna godtas som synonymer utan betydelsedisambiguering. Den första textmeningen i Tabell 55 visar hur i och för sig goda kandidater till synonymer inte kan bytas ut i titlar på verk som i första exemplet (*förebild*). De övriga kandidaterna varierar kraftigt i kvalitet, även om det kan antas att de flesta ordparen med poäng över 4,5, vilka förvisso är få, fungerar bättre. *Folkets synonymlexikon* har poängen 3,0 för ordparet *demokrati – medborgarmakt*, vilket här skulle godtas men hela 4,0 för *politik – tristess*.

När det gäller förslagen från svenska WordNet är den motsvarande svårigheten, som nämnts, att varje ord kan finnas på flera håll i ordnätet och därmed ha skilda, om än för sin respektive betydelse, goda synonymer. Utan en ordbetydelsedisambiguering kan det innebära svårigheter att använda WordNet.

[92] Den information som ordnätet tillhandahåller svarar därmed någorlunda mot inomspråklig (*analytisk*) kunskap, medan textkällan står för den *syntetiska*.

[93] Detta var ett relativt stort arbete, beroende på att lagringsformatet var speciellt. Det verkar vara första gången denna korpus konverterades, åtminstone till Javascript-format. Åke Viberg tackas för att ha gjort undersökningen möjlig, men även Johan Dahl för mycket tillmötesgående när det gällde att praktiskt komma åt texten i det speciella lagringsformatet.

Textmening från Wikipediaartiklar där ord med möjligt synonymiutbyte förekommer	Synonymiförslag från Svenskt OrdNät (WN) respektive Folkets synonymordlista (FS)
karvonen avlade politices doktorsexamen vid åbo akademi 1981 på en avhandling med vårt västra grannland som **förebild** : en undersökning av policydiffusion från sverige till finland .	FS *förebild: föredöme* 4,0*, ideal* 3,5 WN *förebild: förebild, föredöme, dygdemönster*
efter en forskartjänst vid århus universitet i danmark kom han att verka vid åbo akademi och finlands akademi innan han utnämndes til professor i jämförande **politik** vid bergens universitet 1994 .	FS *politik*: *tristess* 4,0
hans forskning har främst rört sig inom jämförande politik och han har även gett ut flera läroböcker kring **demokrati** och statsskick .	FS *demokrati*: *folkstyre* 4,6, *medborgarmakt* 3,0
duane eddy , född 26 april 1938 i corning , new york , är en amerikansk rockmusiker och **gitarrist** .	WN *gitarrist*: *gitarrspelare*
slevar används framför allt till flytande matvaror som **soppa** , **sås** och gryta .	FS *soppa: bensin* 3,4, *oordning* 3,0 FS sås: sky 3,0
slevar görs bland annat i **trä** , plast eller stål .	FS *trä: timmer* 3,0, *virke* 3,0
en soppslev är en större **slev** , ofta med vinklad , djupare , skopa .	FS *slev*: *soppsked* 3,0
l'esprit d'escalier (franska l'esprit de l'escalier ’trappans fyndighet’) är en välfunnen **replik** som man kommer på först när sammanhanget där den skulle fällas inte längre är för handen .	FS *replik*: *besvara* 3,2, *genmäle* 4,3, *svar* 3,4 WN *replik*: *genmäle*

Tabell 55 En redogörelse av synonymiförslag för ord (enbart förekomster i grundform) från *Svenskt OrdNät* (WN) respektive *Folkets Synonymiordlista* (FS) för ord visar att användning av ordutbyten kan vara förenat med komplikationer. De utbytesförslag som anses korrekta har strukits under. Artiklarna är de som förekommer i Tabell 53 och Tabell 54 samt artiklarna *Slev* och *L'esprit d'escalier*. Korrektheten av själva frågorna har ej undersökts. Textformatet är oförskönat, med fetmarkering av ord som har synonymförslag.

En mindre undersökning av hur ofta utbyte till eventuella synonymer av förekommande ord i grundformer gjordes mot Wikipedia-artiklarna *Sahara, Masthugget, Sjö, Kanel* och *Uran*. Här togs den första av eventuellt flera WordNet-ingångar med, samt alla synonymiförslag oavsett poäng från Folkets synonymlexikon. Av träffarna var ett fåtal mycket svårbedömda. De övriga var, inklusive flera synonymiförslag för en enskild ordförekomst i texten 43 stycken förslag från Folkets Synonymordlista och 26 från WordNet. 23 av förslagen från Folkets Synonymordlista (ca 53 %) antogs vara direkt utbytbara, medan 16 från WordNet (ca 62 %) antogs utbytbara. Anledningen till varför förslag inte antogs vara

utbytbara till källornas förslag var i nästan samtliga fall betydelseskillnader men tre förslag var omöjliga att byta ut i kontexten av grammatiska skäl, beroende på annorlunda genus jämfört med den antagna synonymens.

Svenska WordNet har åtminstone använts praktiskt i liknande sammanhang vid ett tillfälle, nämligen av Lönnqvist (2006). Det var där istället fråga om termexpansion vid normal termbaserad dokumentsökning (information retrieval). De olika hyponymerna enligt WordNets SynSet-system hämtades: en sökning efter *bil* skulle expandera själva söktermen till att innehålla *jeep, buss, SUV* osv. Både i tidigare arbeten och här konstaterades att processen att gå från en förekommande term (eventuellt via grundform) till att välja rätt betydelse (*SynSet*) i WordNet är utsatt för mycket svårigheter i form av lexikal ambiguitet. En mycket hög andel av förekommande graford i svensk text är polysema och finns därmed på flera platser i ordnätet. En rent automatiserad användning av hela nätet såsom det beskrivits här ter sig relativt komplicerat, och det finns stor risk för att felaktiga frågor härigenom produceras. Detta är en hållning som intas om systemet som här strävar efter hög precision. I ett längre tidsperspektiv skulle det dock vara möjligt att välja ut bara entydiga termer och begränsa utbytet till synonymer eller hyperonymutbyten till andra termer.

Jämförelser med andra frågebesvarande system med naturligt språk

De jämförelser med andra system som här kommer att göras är haltande beroende på att de andra systemen har byggts under lång period och med en större personalstyrka. Det faktum att systemet självt genererar frågor är också en avgörande skillnad och ett hinder för jämförelser. Allra mest blatant skulle förmodligen en ren jämförelse med det storskaliga och raffinerade projektet *Microsoft PowerSet*[94] (Converse, o.a. 2008) vara. Likheten är dock att det också bygger på Wikipedias databas. Inte desto mindre går det naturligtvis att utan vidare jämförelser påtala de grundläggande skillnader som finns i metodik mellan de två. Det aktuella programmet är byggt för direkt analys av godtycklig text, dvs. det bygger inte på en för-indexering i databas. Den andra stora skillnaden är just att besvarade frågor genereras, dvs. *Microsoft PowerSet* försöker svara på vilken fråga som helst, även de som ej besvaras av databasen (t.ex. *Who is the tallest Dane?,* vilket testats), och levererar konsekvent en rankad listning av de textstycken som bäst antas kunna besvara frågan, *oavsett om informationen verkligen finns eller inte*. En konsekvens är därmed att det inte går att säga på förhand ifall det textsegment som levereras som innehåller ett trovärdigt svar.

[94]Internet-referens, Microsoft PowerSet (PowerSet u.d.)

När det gäller svensk text finns det få system inom området, och de som finns har ganska annorlunda inriktning. Det finns två riktigt tidiga försök, där det första är det nämnda *A natural language parsing program for question answering* (Palme 1971). I rapporten beskrivs programmet SQAP 1. Det andra tidiga projektet, Automatisk textförståelse (ATC), av Welin (1974) beskriver ett program som analyserar text, i källan specifikt nominalfraser. Några utvärderingar av dessa har ej påträffats.

Flycht-Eriksson och Jönsson (2003) redovisar ontologibaserade systemet *BirdQuest* för engelska. En grundläggande skillnad i utvärderingen gentemot här är en kategori, *Out of scope*. Denna kategori innefattar frågor som är omöjliga att besvara för systemet som har kunskap om fåglar, på grundval av att databasen ej täcker just den informationen (*How high does a magpie fly?*), ”socialiseringsfrågor” (*How are you?*) och ännu klarare fall av kunskapsbehov utanför domänen (*How do you kill crows?*). Vidare svårigheter för *BirdQuest*-systemet är de frågor som rättstavade, med annan syntaxkonstruktion eller ordval skulle kunna besvaras av systemet. Dessa utgör 27 % av frågorna. *BirdQuest*-systemet beräknas emellertid i princip klara av att besvara ca 50 % av de frågor som ställs av användare, enligt de rapporterade testerna.

6 Diskussion och framtida forskning

Detta arbete är tänkt att visa att ett fullskaligt parserprojekt kan göras med satsschemat ensamt och utan att bygga på någon av de vanliga träd- eller särdragsbaserade formalismerna som dominerat fältet automatisk analys av svenska, enligt den tillbakablickande uppställningen i kapitel 1. Till skillnad från fullständig parsning innebär satslösning till de funktionella leden på en vald nivå en i hög grad återkommande primärledsmönster per huvudsats. Till skillnad från andra fullskaliga parsrar för svenska överlag fokuserar ansatsen inte lika mycket på uttrycklig matchning av rekursiva led, t.ex. subjektsstrukturer. I denna ansats finns inte en språkdefinierande grammatikkomponent med explicita definierande matchningsregler för de rekursiva strukturerna NP, PP m.m., utan mer heuristiska regler för att känna igen vissa frasgränser för att nå fram till primärledskonstituenterna i Diderichsens satsschema. Om metoden korrekt identifierar de primära begränsade leden innebär det att de områden som kan innehålla rekursiva konstituenter, och därmed viss strukturell ambiguitet, har kapslats in i fält och delfält och ofta kan etiketteras. Ansatsen är öppen för djupare analys, mot fullständig satslösning. Det skulle då ske genom en fortsatt analys av de rekursiva konstituenterna (subjekt, objekt/predikativ och de flesta adverbial) – inklusive satsformade attribut som dessa innehåller. Underordnade satsnivåer har en mer fixerad ordföljd vilket troligen skulle innebära bättre total korrekthet för metoden. Satsförkortningar som infinitivfraser antas kunna analyseras på liknande sätt.

Den aktuella implementationen är som poängterats inte heller riktigt en ”shallow parser” såsom dessa hittills har byggts, med fokus på explicit matchning av flerordskonstituenter som nominalfraser fram till huvudordet. Till skillnad från en ytstrukturell parser finns i ansatsen en central poäng i att klargöra satsnivån hos konstituenterna. För att identifiera de begränsade leden genom bl.a. identifikation av *som*-strykningar används dock även här matchning av flerordskonstituenter (typiskt vissa gränser mellan två konstituenter). Vidare används en mindre specificerad, ’mjukare’, matchning av rekursiva segment för att skilja dessa åt när de är angränsande i samma fält. Matchning av flerordskonstituenter och i så fall vilka, är en avgörande fråga. Istället för att med en grammatik explicit försöka matcha de rekursiva leden inklusive NP och PP, är flerordsmatchningen här till för att underlätta identifikationen av de primära begränsade leden och på så sätt kunna finna de obegränsade leden på annat sätt, helst genom ren uteslutning med satsschemat.

Eftersom termen *parsning* så starkt har kommit att förknippas med en syntaxanalys som sker med hjälp av de grammatikklasser som definierar språket kan det här vara värt att poängtera det ursprung termen har, vilket klargörs av Nivre (2006). Termen parsning som kommer från ett latinskt ord med betydelsen 'dela' har tidigare använts för att beskriva syntaxanalys i en mer allmän mening och det är denna som avses i (induktiv) dependensgrammatisk parsning. På motsvarande grunder beskrivs i föreliggande avhandling en parsning i något allmännare betydelse.

Parsningssystem för svenska har som visats ändå ofta haft den positionsgrammatiska analysen som inspiration och rättesnöre för faktisk grammatikskrivning. Långtgående formella, om än inte helt implementerade, teorier med resonemang angående satsschemats för- och nackdelar för en nordisk grammatikbeskrivning finns framförallt i Lars Ahrenbergs arbeten, t.ex. Ahrenberg (1990).

De allmänna forskningsfrågor om satsschemat som presenterades i inledningskapitlet har fått positiva svar. Det gäller frågan om den positionsgrammatiska beskrivningen är användbar att arbeta efter. Frågan om huruvida den till och med är fördelaktig att utgå ifrån är inte lika lätt att ge ett enkelt svar på. Nedanstående avsnitt ger en bild av vilka skillnader som finns mellan den beskrivna ansatsen och en regelbaserad metod med explicit regelbaserad grammatik för parsning av svensk text.

Struktur i kapitel 6, Diskussion och framtida forskning

I avsnitt 6.1 görs ett försök att karakterisera analysmetoden gentemot de i andra parsrar för svenska som gör annorlunda, ofta mer detaljerade analyser. Detta avsnitt berör också frågan om i vilken grad primärledsmönstren är en ändlig gruppering per huvudsats/finit verbfras, vilket är en relevant fråga för idén att uteslutningsmetoden ska kunna användas.

Avsnitt 6.2 handlar dels om förbättringar i den aktuella analysen som skulle kunna göras för att gagna korrektheten och de tillämpningar som demonstrerats, och dels den möjliga vidareforskning som kan utgå från denna analys.

6.1 Hur ska den heuristiska schemaparsningen jämföras med andra ansatser?

Vad betyder det att automatiskt syntaxanalysera på det sätt som beskrivits i kapitel 2 och 3, utan att använda vad som kanske kan kallas 'standardmodellen för parsning', dvs. en parser med explicit grammatik tillsammans med en parsningsalgoritm, t.ex. en kontextfri grammatik och Earleys algoritm (1970)? En avgö-

rande skillnad är att en på så vis uttryckt grammatikkomponent innebär en uttrycklig *definition* av det i de flesta fall oändliga språket, som i nämnda grammatikklasser. Någon sådan definierande grammatik finns alltså inte med i denna ansats. Det är därmed så att metoden som beskrivs inte kan avgöra grammatikalitet hos en textenhet, eller om textenheten tillhör svenska språket över huvud taget. Algoritmen här kommer alltså att i varje läge försöka analysera också felaktig svenska eller fullständig nonsenstext.

Att metoden helt saknar en explicit definition av språket verkar här dock inte bara vara en svaghet. För det första verkar ansatsen därmed inte antyda att analysatorn, när den parsar, används för att gå ut och ’godkänna’ verklig svensk text gentemot sina regler, vilket det annars kan tolkas som. Det innebär att schemaparsningen är mindre sårbar för variation (”robustare”), och även okänslig för vissa felaktigheter, i syntax, till skillnad från hur parsning med en explicit grammatik som försöker täcka alla möjliga konstruktioner kan vara. Trots att en uttrycklig språkdefinition saknas verkar ansatsen inte ge sämre resultat i korrekthet, utan jämförbara sådana, i de försök till jämförelser som här gjorts (formatskillnader till trots). Som antyds längre ned kan det vara så att en parsning med en explicit grammatik ofta har några allmänna nackdelar som kan mildras hos schemaparsningen här.

Om det antas att några av de systemtyper för parsning av svenska som förekommer skulle kunna förbättras till perfekt eller nära perfekt korrekthet (i förvisso olika detaljerad analys) så antas det föreliggande tillhöra dessa. Det manuella arbetssättet tillsammans med det relativt enkla sätt med vilket förändringar (enkla tillägg eller mer avancerade som tar sig uttryck i helt nya programmerade funktioner med undersökningar av textmeningen under körning) av kod kan ske har givit en positiv grundinställning.

Parsning som redogörelse eller kontroll (träd eller ej)

Den fråga som ska ställas här är vilket syfte parsningen har, en fråga vars varierande svar kan tänkas påverka uppgiften en hel del. I många parsrar för svenska används en trädstrukturell analys. Träd och liknande resultatformat kan sägas ha två roller – dels som ett explicit bevis för att en sträng tillhör en grammatik (*kontroll*), och dels som klargörande av *på precis vilket sätt* den tillhör grammatikens språkbeskrivning (*redogörelse*). I den andra rollen ingår att visa vilka ordsekvenser som utgör vilka fraskategorier (som NP) eller funktionella grammatiska led (som subjekt). Befintliga parsningsprogram fokuserar i väldigt olika grad på de olika rollerna.

I parsrar som används för *grammatikkontroll* är det *kontrollen av att en sträng tillhör språket* som är relevant, och inte redogörelse av analysen. Det är ju bara

relevant att ge alarm för de felaktiga, strängarna som inte täcks på något möjligt sätt. Det kan vara relevant att försöka beskriva felet, t.ex. kongruensbrott i NP, men inte nödvändigtvis att säga exempelvis vilket funktionellt satsled, på vilken satsnivå, en NP utgör.

När empiriska försök görs mot en så god textkälla som publicerad text kan det antas att texten *är* grammatisk, även om det förekommer smärre undantag och skrivfel. Om metoden inte ger täckning för t.ex. SUC kan det vara troligt att själva grammatiken är fel snarare än att språket i texten skulle vara ogrammatiskt. Med andra ord blir uppgiften att kontrollera om textens meningar är korrekta mindre relevant och vad som kvarstår för uppgiften parsning är oftare att ge rätt analys, t.ex. rätt träd, eller som här, etiketterade segment – *redogörelse-uppgiften*.

Ett exempel är hur grammatikkontroll i flera svenska projekt hittills har inbegripit uppgiften ”kontroll av kongruens i NP” för att upptäcka fel som * *en ny recept*. För att göra detta föregås den faktiska undersökningen av korrekt genus/numerus av ett steg där nominalfraser, kongruenta eller ej, identifieras ”grovt” i dessa program. Denna uppgift har en mycket tydlig motsvarighet i denna ansats i den rangbaserade chunkningen, som ju inte undersöker kongruens. Det betyder att chunkningen som används här kan sägas komma från *”kontroll*-aspekten” av parsning – fastän applicerad i en situation där parserns uppgift är *redogörelse*, dvs. att identifiera de funktionella leden, dock inte i trädform. En liknande hållning finns i *CassSwe* (Kokkinakis och Johansson Kokkinakis 1998) där analysens platthet också beskrivs som en fördel. För parsningen här med identifikation som syfte har denna metod visat att undersökning av syntaktisk kongruens mycket sällan behöver ha betydelse för uppgiften. Den segmentering som här sker identifierar därmed *den gamla bordet* (om segmentet antas vara två skilda NP) som en NP, men denna felkategori tillhör inte de kvantitativt viktiga i jakten på korrekthet.

Detta arbete har beskrivit applikationer där *funktionell* grammatisk analys är nödvändig, men där redovisandet av den interna hierarkiska (träd-)strukturen hos fraser/led inte behöver göras. Denna brist på detalj har inte verkat hindra dessa användningsområden. Det är oklart om, eller hur stor betydelse, avsaknaden av själva trädstrukturen skulle behöva ha för andra språkteknologiska områden där funktionell grammatisk analys behövs, t.ex. maskinöversättning. Programmet innehåller inte någon explicit grammatik. Huruvida en positionsgrammatisk modell som satsschemat också bör kallas grammatik är en annan fråga, det är i så fall inte den typ av grammatik som här avses. I någon allmän mening är metoden också ”regelbaserad”, men inte så att själva grammatiken är det på ett språkdefinierande sätt som de den jämförs med här.

Metoden innebär robusthet framför kontroll. Medan en fullständig parsning är mycket väl fungerande för att avgöra om en textmening tillhör det definierade språket, har den heuristiska metod som används enligt ovan inte någon möjlighet alls att göra det. Den antar istället att indata verkligen är en språklig sträng, men försöker (generellt) att parsa vad som helst, och godkänner/förkastar inte uttrycket. Å andra sida gäller en nästan motsatt situation när det gäller att tilldela *korrekt analys*, träd eller ledanalys: den aktuella ansatsen ska ge en enda analys som svar och har denna uppgift som huvudfokus. En fullständig satslösning formulerad med t.ex. kontextfri grammatik kan istället få ett stort antal möjliga svar där uppgiften *att välja rätt analys* med heuristik eller statistik ibland kan vara mycket svår, av olika resultat att döma.

Tydliga skillnader gentemot en parser med en explicit grammatikkomponent

Denna ansats innebär inte någon kritik mot regelbaserade språkdefinierande grammatiker, men det konstateras att användningen av sådana har haft en dominerande ställning för uppgiften syntaxanalys. Flera framgångsrika svenska parsrar har alltså innehållit en explicit grammatikkomponent. Programmet här har däremot inte en sådan grammatik. Ett sätt att granska lämpligheten hos en explicit språkdefinierande grammatik i analyssammanhang är genom att först peka på de erkända fördelarna med dessa, t.ex. en kontextfri grammatik.

Den första fördelen är hur en grammatik uttryckt med omskrivningsregler *exakt* kan redogöra för *om* språket accepterar en sträng eller inte, och att detta kan ske med ett ändligt antal regler fastän antalet möjliga strängar (t.ex. satser) i språket är oändligt stort.

Den andra typiska fördelen med en idealisk grammatik uttryckt med omskrivningsregler är att parsningar av en sträng samtidigt kan åskådliggöra när en sats är möjlig att tolka på flera sätt beroende på strukturell (syntaktisk) ambiguitet. En sats som *Pojken såg flickan med teleskopet* har dubbla tolkningar (beroende på adverbiell eller attributiv tolkning av *med teleskopet*) och kan ges flera möjliga syntaxanalyser vilka kan tydliggöras med t.ex. träd. Satsschemats analys innebär att det nämnda exemplet också får olika analyser medan ambiguitet inuti rekursiva led inte syns.

Dessa två aspekter hos en sådan grammatik pekar kanske samtidigt på svårigheterna med användningen av dem i analys. Komplexiteten för uppgiften är stor som en konsekvens av målsättningen att ge fullständig analys ”på en gång”. I en explicit uttryckt regelformalism riskerar varje ny regel för att täcka en konstruktionstyp att öka antalet möjliga analyser och måste i princip också ackompanjeras med regler som anger precis när regeln ska användas. En felaktig ordklass-

taggning kan i en sådan analys ge komplicerade svåröverskådliga följdfel. Det finns liknande risker hos schemaparsningen, men genom att göra en stegvis analys där begränsade konstituenter relativt framgångsrikt (enligt resultaten här) kan identifieras på viss nivå, och genom att stegen i analysen är mer separerade och inte hänger samman på samma sätt, riskerar metoden inte lika mycket att generera följdfel.

Den andra aspekten, att en sträng kan ge flera analyser med en frasstrukturell, t.ex. kontextfri, grammatik är en mycket stor svårighet som kan göra att parsningsuppgiften *att välja rätt analys* blir en väsentlig del av egentliga arbetet med metoden (avgörandet om strängen över huvud taget tillhör språket är däremot inte svårt). Det rör sig om att koda information om alla fall av naturliga tolkningar som *inte* är flertydiga som exemplet, så att varje grammatikregel enbart används vid rätt tillfällen. Detta är då bara *delvis* en fråga om klara valensregler och i många fall istället en fråga om tendenser och semantiska tolkningsföreträden. (Tolkningen av *av*-PP i *Det hissades en flagga – av plast/av soldaterna* känns i sammanhanget *relativt* lätt att koda rätt med animathetsskalor). Strukturell ambiguitet förekommer långt ifrån bara i formen PP-attachment utan bl.a. i typen ”nivå-ambiguitet” när ett samordnat led ska samordnas på rätt nivå som diskuterades i avsnittet om samordningslicensiering.

Det hävdas inte att den heuristiska positionsgrammatiska stegvisa metoden här kommer undan någon av dessa svårigheter som hänger samman med obegränsade konstituentmönster per sats och strukturell ambiguitet. Även om inte en trädstrukturell beskrivning görs i parsningen, är det så att säga ändå sådant språket är. Men den process för analys som förekommer i denna stegvisa heuristiska analys till vad som kanske kan kallas den minst detaljerade tänkbara funktionellt syntaktiska analysen verkar kunna vara praktiskt användbar. Om förutsättningen, hög korrekthet i identifikationen av de begränsade nyckelkomponenterna har infriats, kapslas rekursiva och därmed potentiellt ambiguösa strukturer delvis in i satsschemats fältdelar.

När det gäller det oändligt stora antalet analyser, innebär valet av ledmönster per sats som redan konstaterats att dessa inte varierar lika kraftigt som mängden möjliga trädanalyser gör för en normal text. Förklaringen är att varje obegränsat (rekursivt) led, med oändligt många möjliga strukturer, abstraheras till bara en variabel, t.ex. primärt subjekt och att analysen här enbart gäller en satsnivå åt gången. Däremot har varje sats (i detta fall är det huvudsatser och primära finita verbfraser som är aktuella) ändå (potentiellt) oändligt många ledmönster. Det beror, som redan nämnts, på att adverbial grammatiskt kan staplas obegränsat och att samordningar av led enligt en tolkning syns i ledmönstret och dessa ledsamordningar kan göras obegränsat. En annan möjlighet till obegränsat antal mönster per *sats* är beroende av hur detaljerad analysen är vid samordningar. En

samordning av ett eller flera led på huvudsatsnivå kan utföras obegränsat antal gånger och om det speglas i ledmönstren leder det till oändligt antal mönster. Med andra ord kan vad som motsvarar en infinit verbfras och som utöver verb innehåller möjliga objekt och adverbial fungera så (*Jag ska köpa en bil, sova, ge dig en båt, stänga dörren eftersom det regnar och gå hem*). Observera att samordning av finita verbfraser inte leder till samma explosion av möjliga mönster per enhet här, eftersom varje primär finit verbfras hanteras separat här, nästan som en huvudsats.

Riktiga svårigheter för analysmetoden påträffas för närvarande i vad som får kallas klara fall av brott mot den moderna grammatiska skriftnormen och klara fall av (strukturell) ambiguitet. Metodens sätt att bygga upp huvudsatser och primära finita verbfraser kring de finita verben leder också till att uttryck med verbellipser för närvarande missas – detta är en svaghet för verbcentrerade ansatser som denna metod. Andra svårigheter som bör vara lika svåra för alla typer av parsrar inbegriper höger-diskontinuerliga relativsatser som modifierar något annat än det som direkt föregår. Generellt är analysen av diskontinuerliga konstituenter inte bra i det aktuella läget, och den rimliga frågan är i vilken grad en fortsatt analys av underordnade nivåer skulle kunna hjälpa till att identifiera fenomenet, vilket logiskt sett borde kunna göras. Satsinskott har också hittills inneburit svårigheter som märks i denna ansats. Sammanfattningsvis får konstateras att det är svårt att hitta konstruktionstyper som är komplicerade att parsa med denna ansats som inte är minst lika svåra med andra grammatikformalismer och parsningsansatser.

6.2 Framtida forskningsfrågor och förbättringar

I fråga om schemaparsningen är programmet som visas i kapitel 4 byggt inte bara som ett gränssnitt för uppvisande av analyser. Syftet är också regelförbättring bl.a. vid åsynen av analyser, t.ex. med olika kriterier eftersöka s-enheter ur träningsmängden bestående av närmare 90 procent av SUC. Det är därför inte helt uppenbart när detta arbete kan anses vara avslutat med en manuell metod som här. Det finns ännu stort utrymme för förbättring och en optimistisk grundsyn när det gäller angreppssättets möjlighet till hög korrekthet. Som nämnts ett flertal gånger skulle själva analysformatet kunna förbättras. Det skulle kunna ske bl.a. genom mer detaljerade kategorier som i *Mamba*-projekten (se Tabell 56 i Appendix). Huvudord identifieras tentativt av programmet och används t.ex. för att göra subjektsidentifikation på grundval av dessa. Denna process skulle kunna göras noggrannare. En stor del av systemets lexikonresurs utgörs av insamlade ordgrupper som märker ut fristående adverbial, mängdord, persontitlar, animata huvudord osv. på ett sätt som enbart ordklasstaggning med särdrag inte gör. Denna ganska stora samling av ordlistor för svenska antas emellertid vara

nödvändig generellt och oberoende av parsningsansats. (Undantag härifrån finns möjligen i system med maskininlärning där handskrivna grammatikregler saknas.) De insamlade listorna är självfallet inte kompletta men skulle kunna utökas med hjälp av extraktionsprogram och/eller manuellt arbete. Suffixanalys skulle kunna få en större betydelse. I vissa fall skulle även suffix kunna svara mot valensinformation och t.ex. ge samma valensingång åt *delbetala* som *betala*.

Som nämnts flera gånger hör också utvidgning av analysen till underordnade satsnivåer till framtida möjligheter. Processen har antagits vara snarlik, med ytterligare omgångar av licensieringsrundor för att avgöra vilka led som finns på sekundär nivå, eller ska licensieras till tertiär nivå osv. De nämnda svåranalyserade enheterna med verbellipser och utbrutna led skulle troligen kunna täckas givet en större tidsram. För de praktiska tillämpningarna som bygger på analysen tillkommer svårigheter som härrör från icke-perfekt ordklasstaggning.

Implementationen som finns har ”webb-egenskaper” men körs lokalt. Det vore möjligt att tillgängliggöra analysen och applikationerna. De skulle kunna göras till webb-applikationer, kanske till server-tjänster och formuleras i kompilerad kod, vilket skulle förbättra hastigheten. Tidsåtgången för analys har generellt haft låg prioritet i detta avhandlingsprojekt.

Ett intressant åtagande för framtiden vore undersökningar rörande statistik för primärledsmönster, och frågan om hur dessa relaterar till textgenre, eller t .o.m. författare. De olika textgenrerna i SUC har inte särbehandlats eller jämförts alls i detta projekt, även om det har blivit tydligt att somliga kategorier som skönlitteratur innehåller en del normbrott som gör dem svårare att syntaxanalysera. I *Mamba*-projekten fanns den potentiella kopplingen mellan språkliga uttryck (transkriberat tal eller skrift) och sociala faktorer som upphovspersonernas bakgrund eller sammanhangets grad av formalitet. Syntaxanalysen i projekten antogs kunna översättas (troligen automatiskt eller halvautomatiskt) till variabler som komplexitet (koncentration), topologi (disposition), specifikation, modalitet och textbindning (Teleman 1974), s. 9–16. Den välformulerade uppställningen från denna källa inbjuder till frågeställningen om en sådan process skulle kunna helautomatiseras.

Parafrasgenerering

Parafrasgenereringen som kan göras genom det beskrivna gränssnittet i kapitel 5 är tänkt att förlita sig på språkkänslan hos användaren och låta denne producera en grammatisk och i övrigt lämplig parafras av en huvudsats. I kapitlet nämndes också andra typer av omformulering som skulle kunna öka möjligheterna och ge en ännu bättre bild av friheten när det gäller innehållsförmedling. Det verkar inte

finnas några direkta användarstudier här i det tidigare modellarbetet av Lindberg och Svensson (1992). Det är inte säkert att funktionaliteten är speciellt intressant för sig själv i en ordbehandlare.

När det gäller denna grundfunktionalitet finns dock intressant forskning som skulle kunna göras. Det vore t.ex. tänkbart att studera hur läsbarhet och förståelse varierar med automatisk syntaktisk omformulering i en text. Det vore möjligt att se vilken effekt texter med t.ex. enbart subjektsinitiala huvudsatser, eller slumpade ledmönster får. Textbindning verkar ju bl.a. manifesteras genom olika spetsställningar. Det vore antagligen möjligt att identifiera olika orsaker till spetsställning av led och därmed försöka identifiera de bakomliggande anledningarna till varje spetsställning. En kvarbliven fråga är den om valfriheten, alltså i vilken utsträckning spetsställning av inget annat syfte än ”variation” förekommer i svensk text. Frågan om vilka begränsningar för spetsställning som råder har ägnats en del nytt intresse från syntaxforskares håll.

Frågegenerering

Applikationen med frågegenerering går i korthet ut på att ge en användare möjlighet att när en svensk text (eller webbsida) öppnas generera frågor som texten besvarar. Denna informationssökningshjälp innebär en funktionalitet som liknar en sökning i texten men där syntaxanalysen och bearbetningen siktar på att möta användarens perspektiv.

Undersökningen om frågegenereringen utifrån en valbar text ställer annorlunda forskningsfrågor jämfört med övriga delar i detta projekt. Applikationen för frågegenerering har absolut flest beroenden: ordklasstaggning, schemaparsning, uppdelning i enkelsatser, spetsställning och slutligen val av frågetyp. Med tanke på detta sena läge i kedjan är den någorlunda höga korrektheten lite överraskande. En hypotes kan vara att text från *Wikipedia* och liknande källor faktiskt är enklare (mer regelbunden) i fråga om syntax. I SUC 2.0 har vissa skönlitterära, juridiska och gammalsvenska stycken, som nämndes, en syntax som är mindre normstyrd och svarar för en stor del av de felexempel som samlats in.

Som beskrivits berör forskningen om denna applikationstyp även vad som är en optimal användarmässig situation. Frågeställningen är hur en användare med en fråga som besvaras av en viss text enklast och snabbast ska kunna finna en formulering som leder till svaret. Framtida arbete härom kan tänkas finna bättre sätt än det här presenterade autokompletteringsgränssnittet genom användarstudier. En annan frågeställning är hur skadliga de frågor som är grammatiskt felaktiga är. Ett lyckosamt utfall vore om dessa ändå vore en hjälp för användare att finna informationen i texten. Det nuvarande programmet har sökegenskaper, dvs. det

formulerar inte om svaret på en fråga utan visar ett antaget relevant textsegment, utan att egentligen riskera att ’ljuga’. Det har också berörts hur den genererade mängden av frågor som per syntaktisk definition har svar kan utökas. Dels är ren omformulering av de befintliga frågorna möjlig så att fler formuleringar görs. Dels är det en spännande uppgift att på andra grunder än rent huvudsatsgrammatiska undersöka extraktion av information som kan uttryckas som frågor i systemet.

7 Summary in English

This Ph.D. thesis describes a method for functional syntactical parsing of Swedish text using the Nordic sentence schema originally introduced for Danish in the 1930's and 40's by Paul Diderichsen (1946). This sentence schema is used here not only as an external template for a rule-based grammar formalism to use. The method instead rather directly seeks to mimic or closely resemble the approach of the manual stepwise exercise of syntactic analysis without an explicit, language-defining grammar component.

	Fund.	**Nexus field**			**Content field**		
Förfält	**Fund-ament**	**Finite verb**	**Subject**	**Sentence adverbial**	**Non-finite verb**	**Object/ pre-dicative, logical subject**	**Other adverbials**
		v	**n**	**a**	**V**	**N**	**A**
Fast *But*	*ni* *you*	*hade* *had*	*[-]* *[-]*	*nog* *probably*	*funnit* *found*	*något nytt* *some-thing new*	*nästa dag.* *the next day.*
Och som nämnts, *And, as mentio-ned,*	*igår* *yesterday*	*hade* *had*	*det* *it*	*faktiskt* *actually*	*kommit* *come*	*en bil* *a car*	*på vägen.* *on the road.*

Table 1 An adaptation of Diderichsen's sentence schema expressed for Swedish functional constituents (*satsled*) on the main clause has roughly seven main positions, plus external positions, like *förfält*. The '[-]' sign indicates a 'trace' – the positional content (in this case the subject) has in this case been fronted to the fundament position from its canonical position. Most full constitituents are eligible for syntactic fronting as well certain subordinate ones producing cleft constructions.

The research questions of this work include whether the sentence schema is a useful description of modern text and if it is feasible or even advantageous to

use it as the core language description for a parser, without having an explicit rule-based definition of the language.

Diderichsen's sentence schema, as shown in Table 1, essentially has a tripartite division of the functional constituents of the clause as nominal, verbal and adverbial (*N, V* and *A*). The approach here has a particular focus on identifying the corner-stones of the clause level examined, in this work the main clause level. These corner-stones are the *bounded* schema position constituents, like *V,* including finite and non-finite verbs but also other ones like particles and reflexive pronouns, see below. In Table 1, the two verb positions are considered bounded. Other bounded constituents, like particles, are also sometimes given a separate schema position in certain versions of the schema. The mid-field adverbials (often of the type *sentence adverbials*) are also to some extent, though not fully, bounded, which still is a useful fact for this approach.

The unbounded functional constituents are the subject, object/predicative and adverbials, whereas the following list shows the important bounded key constituents. Examples come from Stockholm Umeå Corpus 2.0 (Ejerhed, Källgren och Brodda 2006).

- **Finite verbs:** Sedan ***följde*** jag arabens exempel, ***reste*** mig och ***gick*** ut på Aleppos gator. (kk18-119)
- **Non-finite verbs:** Skall ytterligare sprängmassor ***vräkas*** ut? (bb08a-019)
- **Verb particles (always stressed):** Alla föremål strålar ***ut*** värmeenergi. (fh08-035)
- **Reflexive pronouns:** Sen drog han ***sig*** tillbaks hemma i Älvdalen för att vila på lagrarna. (ae01c-030)
- **Sentence conjunctions:** Nå, jag kom dit ***och*** där satt George Kessler och Vreni. (kk10-136)
- **Bounded adverbials:** Staten satsar ***inte*** pengar på lokala trafikleder. (hb06a-069)
- **Förfält (even 'initial extraposition'):** ***En bjässe till resväska,*** det är sant. (ga05-139)
- **Sentence-delimiting markers:** - Kalla honom inte kung i onödan***!*** (kn09-009)

The grammatical output format currently produced is the main clause level (primary level) as shown in the schema, meaning the *full span* of these constituents, including post-modifiers. The analysis is more detailed than a pure sorting of strings into schema positions as it keeps apart multiple constituents in the same position, such as two different objects placed in the *N* position. However, it is less detailed in terms of specifying labelling than both the traditional analysis *('primär satslösning')* and the topmost level marking in the *Mamba* projects from Lund University (Loman och Jörgensen 1971), (Teleman 1974). The difference is e.g. specification of types of adverbials and a number of other categories, and in some versions, marking of head words. These two other forms are still closely related to the schema analysis, in terms of segmentation. Technically, the output format produced by schema parsing here is HTML in the program and an XML format that can be output as in Code example 1.

```
<subjekt>Ni som frågar</subjekt>
<pfv>hade</pfv>
<adverbial>nog</adverbial>
<adverbial>ändå</adverbial>
<piv>kunnat</piv>
<piv>köpa</piv>
<objekt>en vän</objekt>
<objekt>en present</objekt>
<tom>.</tom>
```

Code example 1 The XML output format from the parser for the Swedish sentence '*You, who ask, would anyway probably have been able to buy a friend a present'* includes labels *pfv* (primary finite verb), *piv* (primary non-finite verb) and *tom* (empty).

Identifying the bounded cornerstones on the main clause level at the same time means preparing the task of identifying the instances of unbounded (recursive) functional grammatical categories, namely the nominal subject and object/predicative structures, and the unbounded adverbials. By delimiting the fields on the main clause level (the *primary* level), the restrictions expressed through the schema model mean a starting point in the task of identifying the unbounded constituents through unveiling many secure and uncomplicated conclusions using a method of deduction through elimination and certain formal restriction criteria. A major point of this approach is the frequent possibility to identify *full* functional constituents, like the primary subject, *including* all premodifiers and post-modifiers (phrase-structurally e.g. the maximal projection of an NP), without using an explicit grammar that is applied for matching the corresponding recursive structures. Identifying an entire NP structure, when occurring at the first position on the main clause level, is mostly simply done by identifying the extension of the fundament field, and a simple phrase type identification. An unbounded structure which is not placed in the fundament field or is otherwise delimited by identified bounded constituents and/or end of the clause, can however also be identified, and that likewise without using an explicit grammar, which has to incorporate recursion (since recursive rules are the necessary means for covering unbounded structures). – Identification of full, unbounded constituents in non-delimited spaces obviously demands more from a parser, depending on correctly solving PP-attachment cases and the like regarding the span of a main clause constituent, including post-attributes. The approach, however, still does not use an explicit grammar and can often produce main clause analyses, mitigating difficulties common to 'direct full parsing', such as structural ambiguity within unbounded structures.

Because of coordination of full constituents on the clause level considered, the theoretically unrestricted number of adverbials and possibly a few other phenomena, the set of patterns of sequences of functional categories (like *'subject – finite verb – adverbial – object'*) per main clause is not finite for Swedish. It is, however, the case in practice that these are to a large extent recurring patterns,

with a biased distribution in Swedish text. This means that the set of possible 'results' of a parsed text at the coarse-grained level here is far from as diverse as the set of full analyses (e.g. trees) for the same text. Together with other constraints, identification of functional constituents can partly be carried out by elimination. It is also argued that method being used can just as well be applied for lower clause levels. In fact, the subordinate clause grammar of Nordic languages is more fixed than the main clause level, as regards e.g. subject position, which would probably lead to an increase in total correctness. As this work is concerned with the primary clause level, it does not provide an analysis of the internal structure of these subordinate clauses, which constitute or belong to unbounded main clause constituents.

A consequence of the lack of explicit phrase grammar is that the method thereby has fewer rules involving matching multi-word units, like NPs, PPs and other unbounded structures. This is seen as an advantage as explicit matching rules of such sequences need to be very many for good coverage and a long segment often corresponds to several structures in a grammar (structural ambiguity). The current approach, however, is not free from matching of multi-word units. In fact, the rules initially identifying the primary bounded constituents, which in a sense are the prerequisite for identification without multi-word units, themselves include these kinds of rule patterns, e.g. in identification of *that*-deletion (see below) which is part of the identification of primary bounded constituents. Explicit matching of multi-word units is also applied when identifying *förfält* segments and certain adverbials.

The Stockholm Umeå Corpus (Ejerhed, Källgren och Wennstedt, o.a. 1992)*, version 2.0* (Ejerhed, Källgren och Brodda 2006), a genre balanced corpus mirroring proportions of modern published Swedish text of a million words has been used throughout this work. The role of the corpus is to represent Swedish text and to provide correctly POS-tagged texts, so that this project can focus on the parsing task. The corpus is also used for building a parser, making the program able to work with unrestricted text. The corpus has been randomly divided up into two sets of s-units (s-units are sentences and other units like headings etc.), where about 89 % are in the training set used for manual iterative parser improvement and the rest, the test set, is saved for evaluations.

Identification of bounded constituents

In a similar manner to the manual schoolbook exercise of syntactic analysis of Swedish, the technique here starts off by identification of the finite verbs, which are considered compulsory and have fixed positions, 'V2'. Also in accordance with the manual exercise, it is the finite verbs *on the main clause level* which are

searched for. (As mentioned, the current work is devoted specifically to this functional main clause analysis.) The act of identifying these and the other present bounded *primary* (main clause level) constituents has particularly important implications in the current parsing method, which thereby delimits main clauses, primary verb phrases, schema fields and also sometimes the various positions therein. The techniques used for identification of the different types of bounded constituents are similar, and are here expressed as for finite verbs, which, as mentioned, are the compulsory elements in the clause.[95]

The general approach for identification of primary constituents is a process of collection of candidates followed by a *licensing* procedure. In this case, licensing means taking away candidates for primary units, through identification of non-primary sequences like sub clauses and reported segments. In this approach, following *Svenska Akademiens grammatik* (Teleman, Hellberg och Andersson 1999), a reporting construction like '*It was cold, we said'* interprets *it was cold* as a primary object, and therefore as a subordinate clause. As for verbs, the tagging information makes it easy to collect the candidates: finite verbs are either in present or preterit tense or imperative. The licensing procedure consists of several passes, identifying clear cases of complementizers (*eftersom/since, var/where, vilket/which*), reasonably clearly corresponding to a group of POS tags in the tagset of SUC 2.0.

Left boundaries of subordinate clause structures are also found through identification of 'covert complementizers', *that*-deletion. In Swedish, the word *that* (*who, which* etc.) corresponds to *som* or to the complementizer *att*. The two 'invisible' words can be identified through a collection of multi-word patterns which look a bit different in the two cases. *That*-deletion where the covert word is *som* is often in the form of two adjacent nominal structures (such as *the man I*/ *mannen jag*), where the second structure is the subject of the relative clause and may be case-marked as subject. *That*-deletion where *att* is missing often takes on the form of two rather close finite verbs, where the first one is a reporting verb (such as *thinks it works/tycker det funkar, said I we did/sa jag vi gjorde*). Currently, about 50 patterns are used for detection of *that*-deletion. Positions of starts of sub clauses (through overt or covert complementizers) are stored. Passes from left to right keep track of how many such licensors are encountered. This can be expressed as a stack structure, since sub clauses are often nested (*We think <u>that</u> the house <u>that</u> we <u>saw</u> <u>was</u> new*). When a finite verb (if that is the

[95] There are exceptions to many of the 'general' statements regarding Swedish grammar. As for finite verbs in Swedish main clauses, an exception of the idea is primary *kanske*-clauses (*maybe*-clauses), *Han kanske gått, (He maybe gone)* – clearly because of the historical origin of *maybe* as a verbal complex (Andréasson 2002), which, however, is normally tagged as an adverbial.

bounded constituent to be identified then) is encountered with a non-empty stack, the verb is normally licensed – it is no longer considered a candidate, and one licensor is popped off the stack.

The next way of licensing bounded constituents is identification of V1-conditionals (*Vore det inte för regnet skulle jag komma/Were it not for the rain, I would come*). These constructions can be discovered using a combination of different matching rules. What is important here, is to separate these types of constructions from other V1-formed constructions as *yes/no*-questions, imperatives and 'diary-style' (*Went to the store, bought coffee*). One important rule, obviously, is that the verb considered to be primary should be followed by a potential subject.

Licensing by coordination is a special kind of licensing that applies to normal (not '*pairwise'* like *antingen/either*) conjunctions directly followed by a finite verb. In a case like *Om de som besökte och festade åkte skulle vi följa med/If those who visited and partied left, we would follow,* there are two licensors, but there are three subordinate verbs (*besökte/visited, festade/partied, åkte/left*). The method here is not to license a verb directly following that type of conjunction but to leave *festade/partied* for later processing, licensing by coordination, which in clear cases like this should be coordinated with a verb on a subordinate level.

In relatively few cases, the rules for e.g. detection of *that*-deletion, which must be carefully formulated not to be triggered erroneously, fail to license all bounded constituents that should be licensed in a way that is obvious and discovered by the parser. As a rule, only one primary finite verb is to be found in each main clause or coordinated primary finite verb phrase. It is around these finit verbs that the main clauses and finite verb phrases are formed. If, however, more than one finite verb remains unlicensed in a segment without any possible coordinator in between, one of the verbs should be licensed. In this case, the program works by using similar licensing techniques but with a heuristic rule relaxation. A typical case is a V1 conditional in something other than the first main clause: *Det finns inget recept, men uppträder man korrekt hoppas jag att det smittar av sig.*/Almost literary: *There are no recipes, but acts one correctly (if one acts correctly) I hope it will spread*. In this case *uppträder/acts* and *hoppas/hope* are left unlicensed after the normal licensing process. *Uppträder/acts* is in the program currently licensed by late detection of V1-conditional – although directly following the coordinator *men/but*.

Accuracy of identification of the bounded primary constituents is generally high. Primary finite verb identification is estimated to have an f-score of about 98 % in Wilhelmsson (2008).

Identification of unbounded constituents

When the bounded primary constituents are identified, main clauses and primary finite verb phrases are thereby also identified and split up further by the optional bounded constituents. This situation provides a starting point for the identification of unbounded constituents where the sentence schema often allows only one primary functional constituent in a delimited schema section. A particularly favourable syntactic construction is the primary auxiliary verb construction, which creates a delimited *midfield* (*nexus field*). In this type of clause the subject can mostly be easily identified using the simple logic of the sentence schema. For this reason, an experimental expansion of the class of auxiliary verbs is tested. Since the level of analysis here means a restricted number of nominal constituents per main clause or coordinated finite verb phrase, each such certain identification of a constituent also helps the rest of the analysis, by elimination.

Wordclass/word	**Example**	**Rank**
Som tagged as conjunction	*Som målvakt var han bra.*	16
Preposition	*Till, för*	15
Word in genitive case	*Kalles, bokens*	1/14
Determinator	*De, några*	5
Possessive	*Dess, sitt*	4
Cardinal number	*43*	3
Adverb	*Ganska, bra, bort*	3
Participle	*Slående*	2
Adjective	*Grön, hoppfulla*	2
Ordinal number	*43:e, första*	2
Measure noun	*Kopp, kilo, handfull, msk*	1.5
Personal title	*Herr, cupvinnaren, tvåan*	1.5
Personal pronoun	*Han, de*	1
Proper name	*Karl, Karlsson, Paris*	1
Noun	*Idé, elefanter*	1

Table 2 By assigning each word in the non-analyzed areas a rank according to this table, many boundaries between phrases like NPs, PPs, *som*-phrases, adverb phrases, adjective phrases and other segments can be recognized. As is seen, few POS feature values other than the word class is used, which often mitigates the effect of errors in the POS tagging. The categories with a rank of 1.5 are usually nouns (rank 1) but have possible NP complements. These groups consist of manually produced listings of words.

It is, however, not always the case that unbounded constituents can be clearly identified without having to examine the structure at all. When structures occur at other positions than in the fundament, there is a need for identification (finding the span of an unbounded segment) and initially labeling it in terms of

phrase type, such as NPs, PPs, other phrases and segment types. The fields not occupied by the primary bounded constituents are examined using a specialized technique for this segmentation using the particular POS tagset of the *Stockholm Umeå Corpus*. This technique, *rank-based chunking*, assigns a fixed rank to each of the words in these areas, see Table 2, and uses those to identify 'minimal' NPs (and AdvP, AdjP, participle phrases and numeral phrases), PPs and *som*-phrases – this can be seen as NPs and some others, without post-modifiers, possibly after a preposition, possibly after a *som*, which is marked as a conjunction, see Table 2.

After assigning ranks to the words, first pass assigns boundaries between the chunk segments, and the types of these can be decided by simple rules.

1. The pass from left to right works in the following way. Any rank number is supposed to have a lower number *('higher rank'*) than previous word, or the same rank, to be considered as part of the same chunk.

   - If the rank is lower (higher number), the interpretation is that the previous chunk is terminated a new one is started.
   - The same goes for adjacent two words of the rank 1, unless those are forming a personal name (a list of 21 000 personal names is currently used, thanks to Dimitrios Kokkinakis).

2. The type of a chunk can be established by looking mostly at the first word, or the head word (which in most cases is the one of highest rank, in the end).

   - If the first word is a preposition (rank 15), it is a PP
   - If the first word is *som* (conjunction, rank 16), it is a *som*-phrase
   - In other cases the structure is something else, in most cases nominal: NP, adverb phrase, adjective phrase

This chunking step can be said to be rather cautious and prefers postponing other decisions about merging segments than those directly inferred by the algorithm. After this chunking step, segments are merged further aiming at a final number of nominal segments compatible with the clause and verb type. This further segmentation is carried out in three layers.[96] The steps thereby also produce other types of structures, like infinitival phrases and sub clauses. The steps include merging motivated by valency for other word classes than verbs, (either manually stated or extracted from databases used for lexicon production: *Natio-*

[96] There is no particular reason for using exactly three layers, other that their display helps localizing errors etc. The layers do not represent phrase structure or anything similar.

nalencyklopedins ordbok (1995–96) and *Lexin – svenska ord* (1998)), heuristic matching of sub-clauses and default merging of the whole fundament to one constituent. The program has been built to help adding links between the chunks through the graphical interface.

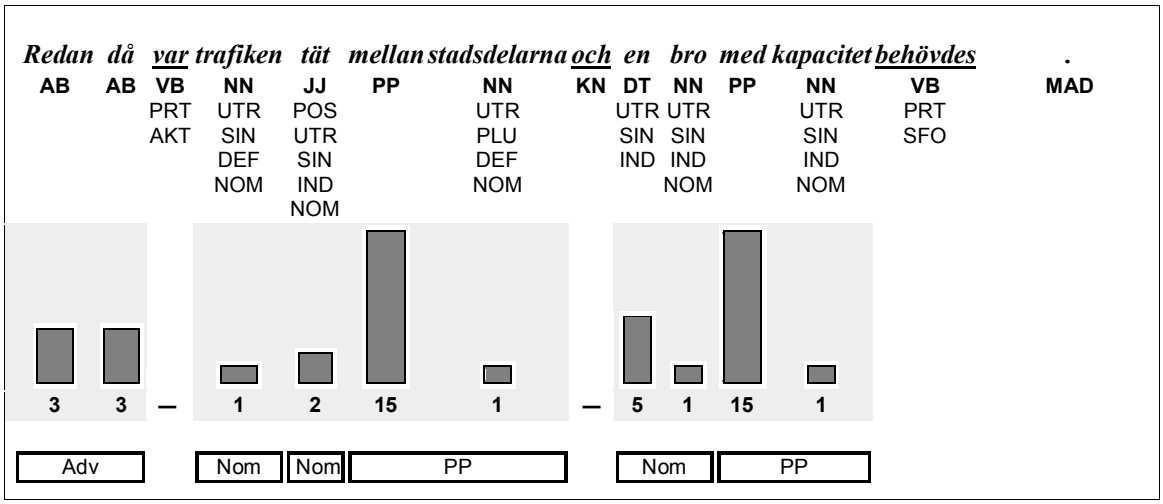


Figure 1 The rank-based chunking produces nominal and adverbial chunks as shown at the bottom using the ranks, illustrated with the height of bars (hb06a-011). The meaning of the POS tags is explained in Tabell 6 in chapter 2. Translation: *Already then, the traffic between the parts of the city was heavy, and a bridge of some capacity was needed.*

After these steps, the potential candidates for subject and object/predicative are collected in each main clause and coordinated primary finite verb phrase. Whereas certain phrase types, such as PPs, AdvPs, when identified separately, have rather clear corresponding functional roles (namely as adverbials), many NPs with certain head words may function as adverbials instead of subjects/objects/predicatives. The labelling step that finalizes the parsing is carried out using a number of different sub functions for different clause and VP types. As mentioned, auxiliary clauses are relatively easy to analyze, whereas correct labelling in other transitive sentences has attracted attention from researchers, e.g. Øvrelid (2008), explaining default interpretations in clauses with potential subject/object ambiguity, partly through prominence scales involving both formal and semantic aspects.[97] A sort of prominence ranks is used in the parser, involving manual listings to identify e.g. human subject candidates. Also, the subject identification involves several special cases for certain verbs, e.g. copulas.

[97] *Brevet (OBJ) fick flickan (SUB),* lit. *The letter got the girl* is correctly analyzed since *the girl* is human and thus has a higher position on the animacy scale as compared to *the letter*.

As for the results in correctness, it should probably be stressed that this type of work with manual rule improvement by investigating errors looking at a large training subset of the corpus shows a current state of the parser accuracy. Limitations come in the form of restricted time for work with this type of 'bug fixing', and almost never because the correct analysis seems impossible to produce (unless the POS tagging is wrong), using a free, very flexible and expressive object based representation of the sentence under analysis. The implementation in this work roughly mirrors the stepwise approach in terms of modularity and particular sub functions used for certain syntactic constructions.

The correctness of this parsing method is high, although the manual improvement method through bug-fixing used seems to have much better potential, given more time. The evaluation here has been carried out with a correctly tagged test set from SUC 2.0. F-score for primary finite verb identification has been evaluated to 98 %. Exact matching of primary subjects (including postmodifiers) has shown to be correct in more than 90 % of all main clauses. Marking of other primary nominal constituents as well as primary adverbials is correct in at least 70 % of the cases.

This method for syntactic analysis produces an output form which is not exactly the same as that of any other parser for Swedish. Compared to shallow parsers (whether current system is considered as a shallow parser or not), the difference is that the clause level is distinguished so that explicit matching of recursive constituents is partly avoided. Several parsers are built for full analysis 'directly'. This is obviously a more detailed analysis, and the current parsing output is in some cases, like when compared to the dependency parsing projects *MaltParser* (Nivre 2006) and *SweFDG* (Voutilainen 2001) clearly a derivable, less detailed form of functional analysis, although a future stepwise expansion using schema parsing could also reach a not too different form.

Applications built using heuristic schema parsing

Having an 'intermediate' level of analysis in the output format – maybe the least detailed functional analysis possible – it is examined how this type of format can be used in language technology applications for unrestricted Swedish text. The two pilot applications developed are both new kinds with few similar counterparts in research. In fact they are dependent on having at least this detail of (functional grammatical) analysis. For these prototypes, a part-of-speech tagger was built with an average correctness, around 95 % in estimated correctness. The tagger, which could be improved or replaced, guarantees a reasonably good input format, and should give a realistic or worse view of the practical usability.

The first application examines the possibility of an automatic, or more specifically, user-defined mechanism for syntactic fronting, in accordance with the rules of Diderichsen's sentence schema, in simple main clauses. The graphical interface has an editor and shows the analysis of written sentences in a way so that different (unbounded) functional constituents can be clicked in the interface and thereby fronted, and inserted into the original text. This type of application thus leaves the final decision regarding the appropriateness of a certain constructions to the user, similarly to other editorial tools. But it is also discussed what the actual limitations are in fully automatic paraphrasing of Swedish text, as well as what is natural. In itself, the application for syntactic fronting could be used for examining these aspects involving such areas as text cohesion, truth preservation and text comprehension. This application has not been evaluated and does not incorporate all of the restrictions for syntactic fronting that are discussed. It seems already able, however, to be used for experimentally answering questions such as 'What happens if all subjects are fronted in a text?', or, 'What happens if syntactic fronting is carried out completely randomly in a text?' In a wider perspective, the functionality could also be seen as part of the research in grammatical naturalness, and could perhaps be helpful in areas like text generation.

The last application in this work is a program that belongs to the area of natural language query systems. The application has an input text as its database, in the current application either an article from the downloaded database of Swedish *Wikipedia*, or other text input. The application provides the user with the possibility of posing questions regarding the text, though restricts what questions the user can ask. A system might otherwise run the risk of giving erroneous or only 'best match' answers. This type of application lets the user decide or be aware of what type of data base is available, clearly restricting the potential queries from the user. Instead of a pattern matching approach, the system tries to explicitly generate the questions that are answered by the text, essentially by grammatical definition, see below. The set of questions available to the user is roughly the number of primary unbounded functional constituents in the text. In this case, the user interface has been built to show the actual text from the start and making use of an auto complete drop-down list for restricting the user to available questions. Whether this is an optimal way of leading the user to an available formulation of a question in mind is not clear, and this question of user interaction could be examined further. The actual answering of the question is done by highlighting the sentence from which the question was produced, and scrolling it into view so that a few earlier sentences are also visible to the user so that she may solve possible anaphoric references herself.

The idea here has been to keep the precision value (the correctness of the answers, given the questions) as high as possible. The program tries not to take

risks by producing potentially unanswered questions. The negative side would then naturally be the recall (the number of questions produced as related to the number of all possible questions, or even *formulations of questions*, that the text information can be said to provide answers to). It is discussed whether this set could be expanded through use of synonymous expressions, extraction of propositions from subordinate clause levels or even more advanced techniques, like finding e.g. inter-sententially stored information.

The actual question generation technically builds on the syntactic fronting technique. In main clauses and primary finite verb phrases – which are turned into main clauses by inheriting the subject from previous main clauses – all versions of different fronting are produced. *Idag köpte vi inte några nya mattor* (*Today, we did not buy any new carpets*) thus produces versions with fronted adverbials, subject and object. Each such fronted reading can possibly be turned into a *wh*-question by substituting a *wh*-phrase for the fronted constituent. This is of course not obviously possible for all constituents like the sentence adverbial here (*inte/not*). In the question generation for nominal constituents (subject and object/predicative), there is a rather small set of *wh*-words (*what, who, which*) that should be possible to choose among by examining head words. As for the adverbials, the situation is more complex. As for adverbials formed from adverb phrases, participle phrases and NPs, there should exist a reasonably clear, though vast and not fully covered mapping to corresponding *wh*-words. As for prepositional phrases, the correct mapping functionality to *wh*-words must rely sometimes on the head word (the preposition), but for many prepositional phrases (like the common ones starting with *i, på, till*), also the head of the complement, or the whole complement. A much simpler form of question generation is to generate *pied-piping* versions instead, but it is believed those questions, though correct, are less valuable in actual use of the application since they appear to require more knowledge regarding the actual form of the information from the user. A tentative mapping functionality grouped according to the word classes appearing in SUC 2.0 is presented.

The results of the current prototype of the question generation shows that only the actual choice of *wh*-words often is what goes wrong. This is a positive result since the mapping to *wh*-words is very tentative and can be greatly improved with the same methodology. In a minor test, 64 % of the produced questions were considered entirely correct, whereas most of the errors seem possible to avoid given more time, e.g. by classification of the head words of PP complements.

Referenser

”A Dictionary of Philosophical Terms and Names.” http://www.philosophypages.com (använd den 01 12 2009).

Abney, Steven. ”Parsing by Chunks.” i *Principle-Based Parsing*. Kluwer Academic Publishers, 1991.

Abney, Steven. ”Part-of-Speech Tagging and Partial Parsing.” i *Corpus-Based Methods in Language and Speech Processing*, av S Young och G BlootHooft, 118-136. Kluwer Acad. Publishers, 1997.

Ahrenberg, Lars. ”A grammar combining phrase structure and field structure.” *Proceedings of the 13th conference on Computational linguistics - Volume 2*. Helsingfors: Association for Computational Linguistics, 1990. 1 - 6.

Akademien, Svenska. *Svenska Akademiens ordlista (SAOL), 11:e upplagan.* Stockholm: Norstedts Förlag, 1986.

Allén, Sture. *Tiotusen i topp: ordfrekvenser i tidningstext.* Stockholm: A&W, 1972.

Andersson, Erik. *Grammatik från grunden: en koncentrerad svensk satslära.* Uppsala: Hallgren & Fallgren, 1994.

—. ”Tidsadverbial och syntaktisk struktur.” *Förhandlingar vid sammankomst för att dryfta frågor rörande svenskans beskrivning 8.* Lund: Lunds universitet, Christer Platzack, 1974. 1-14.

Andréasson, Maia. *Kanske - en vilde i satsschemat.* Miss 41, Meddelanden från Institutionen för svenska språket vid Göteborgs universitet, 2002.

—. *Satsadverbial, ledföljd och informationsdynamik i svenskan (Doktorsavhandling).* Göteborg: Institutionen för svenska språket, Göteborgs universitet, 2007.

Bigert, Johnny, Ola Knutsson, och Jonas Sjöbergh. ”Automatic Evaluation of Robustness and Degradation in Tagging and Parsing.” *In proceedings of RANLP 2003 (Recent Advances in Natural Language Processing).* Bovorets, Bulgarien, 2003.

Birn, Juhani. ”Diskontinuerlighet som strukturenlig företeelse.” *Svenskans beskrivning 18, 1990.* Uppsala: Lund Univerity Press, 1991. 75-86.

—. *Swedish Constraint Grammar: A Short Presentation.* 1998. http://www2.lingsoft.fi/doc/swecg/intro/.

Borin, Lars, Markus Forsberg, och Lennart Lönngren. ”The hunting of the BLARK - SALDO, a freely available lexical database for Swedish language technology.” i *Resourceful language technology. Festschrift in honor of Anna Sågvall Hein*, av Joakim Nivre, Mats Dahllöf och Beáta Megyesi, 21-32. Uppsala: Acta Universitatis Upsaliensis: Studia Linguistica Upsaliensia 7, 2008.

Brants, Thorsten. ”TnT - A Statistical Part-of-Speech Tagger.” *Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000.* Seattle, WA, 2000.

Bresnan, Joan. *Lexical-Functional Syntax.* Oxford: Blackwell Publishers Ltd, 2001.

Brodda, Benny. *(K)overta kasus i svenskan.* Stockholm, doktorsavhandling: Stockholms universitet, PILUS 18, 1973.

Börjars, Kersti, Engdahl Elisabet, och Maia Andréasson. ”Subject and object positions in Swedish.” *Proceedings of the LFG03 Conference.* 2003.

Carlberger, Johan, och Viggo Kann. ”Implementing an Efficient Part-Of-Speech Tagger.” *Software—Practice & Experience*, 1999: 815 - 832.

Converse, Tim, Ronald M Kaplan, Barney Pell, Scott Prevost, Lorenzo Thione, och Chad Walters. ”Powerset’s Natural Language Wikipedia Search Engine.” *Wikipedia and Artificial Intelligence: An Evolving Synergy. Papers from the 2008 AAAI Workshop.* Chicago, USA: AAAI Press, 2008. 67.

Diderichsen, Paul. *Elementær Dansk Grammatik.* Köpenhamn: Gyldendahl, 1946.

—. *Helhed og struktur, utvaldte sprogvidenskabelige afhandlinger.* Gad, 1966.

Diderichsen, Paul. ”Sætningsbygningen i Skaanske Lov — Fremstillet som Grundlag for en rationel dansk Syntaks.” Doktorsavhandling, Köpenhamn, 1941.

Earley, Jay. ”An efficient context-free parsing algorithm.” *Communications of the Association for Computing Machinery 13:2*, 1970: 94-102.

Einarsson, Jan. *Talbankens skriftspråkskonkordans.* Lunds universitet: Institutionen för nordiska språk, 1976.

Ejerhed, Eva. ”En ytstrukturgrammatik för svenska.” i *Förhandlingar vid Sammankomsten för att dryfta frågor rörande svenskans beskrivning*, av Sture Allén, Lars-Gunnar Andersson, Jonas Löfström, Kerstin Nordenstam och Bo Ralph. Göteborg, 1985.

Ejerhed, Eva, Gunnel Källgren, och Benny Brodda. *Stockholm-Umeå corpus version 2.0.* Institutionen för Lingvistik, Stockholms universitet, Institutionen för Lingvistik, Umeå universitet, 2006.

Ejerhed, Eva, Gunnel Källgren, Ola Wennstedt, och Magnus Åström. *The Linguistic Annotation System of the Stockholm-Umeå Corpus Project.* Rapport no. 33, Umeå Universitet: Instititutionen för lingvsitik, 1992.

Engdahl, Elisabet. ”'Det var det ingen som ville': om spetsställningens form funktion och restriktioner i svenska.” *Svenskans beskrivning 23.* Lund, 1999. 96-105.

Engdahl, Elisabet, och Eva Ejerhed. *Readings on Unbounded Dependencies in Scandinavian Languages.* Umeå: Almqvist & Wiksell International, 1982.

Faarlund, Jan Terje, Svein Lie, och Kjell Ivar Vannebo. *Norsk referansegrammatikk.* Oslo: Universitetsforlaget, 1997.

Flycht-Eriksson, Annika, och Arne Jönsson. ”Some Empirical Findings on Dialogue Management and Domain Ontologies in Dialogue System - Implications from an Evaluation of BirdQuest.” *Proceedings of 4th SIGdial Workshop on Discourse and Dialogue.* Sapporo, Japan, 2003.

Forsbom, Eva. ”Good Tag Hunting: Tagability of Granska Tags.” i *Resourceful Language Technology: Festschrift in Honor of Anna Sågvall Hein, ACTA UNIVERSITATIS UPSALIENSIS Studia Linguistica Upsaliensia 7*, av Joakim Nivre, Mats Dahllöf och Beáta Megyesi. Uppsala, 2008.

Gambäck, Björn. *Processing Swedish Sentences: A Unification-Based Grammar and some Applications (doktorsavhandling).* Stockholm: KTH och Stockholms universitet, 1997.

Göteborgs, Svenska akademien och Institutionen för svenska språket vid. *Svensk ordbok.* Norstedts, 2009.

Hellberg, Staffan. ”Varför inte prepositionsobjekt?” i *Grammatik i fokus. Festskrift till Christer Platzack*, av Lars-Olof Delsing, Cecilia Falk, Gunlög Josefsson och Halldór Sigurðsson, 47–53. Lund: Institutionen för nordiska språk, Lunds universitet, 2003.

Henriksen, Carol. ”Sætningsleddene og deres stilling - nogle år før - og flere år efter.” *NyS : Nydanske studier og almen kommunikationsteori, 16/17: Saetningsskemaet og dets stilling - 50 år efter*, 1986: 210 - 228.

Hobbs, Jerry, o.a. ”FASTUS: A Cascaded Finite-State Transducer for Extracting Information from Natural Language Text.” i *Finite State Devices for Natural Language Processing*, av Emmanuel Roche och Yves Schabes. Cambridge, MA: MIT Press, 1996.

Holm, Gösta. ”Svenska Akademiens grammatik (recension).” *Arkiv för nordisk filologi 115*, 2000.

Holm, Lars, och Kent Larsson. *Svenska meningar: Elementär språklära.* Lund: Studentlitteratur, 1980.

Hultman, Tor G. *Svenska akademiens språklära.* Stockholm: Svenska akademien. Norstedts ordbok distributör, 2003.

Jespersen, Otto. *The Philosophy of Grammar.* London: Allen & Unwin, 1924.

Johansson Kokkinakis, Sofie. *En studie över påverkande faktorer i ordklasstaggning. Baserad på taggning av svensk text med EPOS.* Doktorsavhandling, Göteborg: Institutionen för svenska språket, Göteborgs universitet, 2003.

Josefsson, Gunlög. *Svensk universitetsgrammatik för nybörjare - Övningar med facit och kommentarer.* Lund: Studentlitteratur, 2001.

—. *Svensk universitetsgrammatik för nybörjare.* Lund: Studentlitteratur, 2001.

Järborg, Jerker. *Manual för syntaggning [Manual for syntagging].* Göteborgs universitet: Institutionen för språkvetenskaplig databehandling, 1986.

Järborg, Jerker, och Pernilla Danielsson. ”The PAROLE report (WP-4.2.2b) "Morphosyntactic Description of Swedish".” 1996.

Jörgensen, Nils, och Jan Svensson. *Nusvensk grammatik.* Malmö: Gleerups, 1986.

Kann, Viggo, och Magnus Rosell. ”Free Construction of a Free Swedish Dictionary of Synonyms.” *Proceedings of 15th Nordic Conference on Computational Linguistics – (NODALIDA 05).* Joensuu, 2005.

Karlsson, Fred, Atro Voutilainen, Juha Heikkilä, och Arto Anttila. *Constraint Grammar - A Language-Independent System for Parsing Unrestricted Text.* Berlin och New York: Mouton de Gruyter, 1995.

Knutsson, Ola. *Developing and Evaluating Language Tools for Writers and Learners of Swedish.* Stockholm: Doktorsavhandling, KTH, 2005.

Knutsson, Ola, Johhny Bigert, och Viggo Kann. ”A Robust Shallow Parser for Swedish.” *Nodalida 2003.* Reykjavik, Island, 2003.

Kokkinakis, Dimitrios. ”More than Surface-Based Parsing; Higher Level Evaluation of Cass-SWE.” *13th Nordic Computational Linguistics Conference (NODALIDA)*. Uppsala, 2001.

Kokkinakis, Dimitrios, och Sofie Johansson Kokkinakis. *A Cascaded Finite-State Parser for Syntactic Analysis of Swedish.* Research Reports from the Department of Swedish, Göteborg: Institutionen för svenska språket, Göteborgs universitet, 1998.

Kvist Darnell, Ulrika. *Pseudosamordningar i svenska: särskilt sådana med verben sitta, ligga och stå (Doktorsavhandling)*. Stockholms universitet: Institutionen för lingvistik, 2008.

Källgren, Gunnel. *En algoritm för bestämning av subjekt och direkt och indirekt objekt i löpande text.* Institutionen för lingvistik, Stockholms universitet, 1989.

Källgren, Gunnel. *Making maximal use of surface criteria in large-scale parsing : the MorP parser.* Papers from the Institute of Linguistics, University of Stockholm, (PILUS) 60, Stockholm: Institutionen för lingvistik, Stockholms universitet, 1992.

Landqvist, Hans. *Författningssvenska - Strukturer i nutida svensk lagtext i Sverige och Finland, Doktorsavhandling.* Göteborgs universitet, 2000.

Lexin – Svenska ord. Norstedts Akademiska Förlag, 1998.

Lindberg, Janne, och Carin Svensson. *Topikalisering som skrivstöd. En implementering med satsschema.* Rapport från IPLab-projektet, Interaktions- och Presentationslaboratoriet, Språkliga datorstöd vid skrivande 8, TRITA-NA-P9225, IPLab-60, Stockholm: Tekniska högskolan i Stockholm, Stockholms universitet, Numerisk analys och datalogi, 1992.

Ljung, Magnus, och Sölve Ohlander. *Allmän grammatik.* Malmö: Gleerups, 1971.

Loman, Bengt, och Nils Jörgensen. *Manual för analys och beskrivning av makrosyntagmer.* Lundastudier i nordisk språkvetenskap, Serie C nr 1, Lund: Studentlitteratur, 1971.

Lundqvist, Aina. *Språklig anpassning - Syntaktisk analys av ett barnboksmaterial (Doktorsavhandling)*. Göteborg: Institutionen för svenska språket, Göteborgs universitet, 1992.

Lönnqvist, Susanna. *Expansion av sökfrågor med Svenskt OrdNät som termkälla.* Borås: Magisteruppsats i biblioteks- och informationsvetenskap, Högskolan i Borås, 2006.

Megyesi, Beáta. *Data-driven Syntactic Analysis: Methods and Applications for Swedish.* Doktorsavhandling, Stockholm: Institutionen för Tal, Musik och Hörsel, KTH, 2002.

Melin, Lars, och Sven Lange. *Att analysera text - Stilanalys med exempel.* Lund: Studentlitteratur, 1986.

Nationalencyklopedins ordbok. Höganäs: Bra Böcker, 1995–96.

Nivre, Joakim. *Inductive Dependency Parsing.* Dordrecht: Springer, Text, speech, and language technology series, Volume 34, 2006.

Nivre, Joakim, Leif Grönqvist, Malin Gustafsson, Torbjörn Lager, och Sylvana Sofkova. ”Tagging spoken language using written language statistics.” *International Conference On Computational Linguistics, Proceedings of the 16th conference on Computational linguistics - Volume 2, Köpenhamn.* Morristown, NJ, USA: Association for Computational Linguistics, 1996. 1078-1081.

Nivre, Joakim, o.a. ”MaltParser: A language-independent system for data-driven dependency parsing.” *Natural Language Engineering, 13(2)*, 2007: 95-135.

Palme, Jacob. *A natural language parsing program for question answering.* FOA P rapport, Stockholm: Försvarets forskningsanstalt, Planeringsbyrån, 1971.

Palme, Jacob. *A Natural Language Parsing Program for Question Answering.* FOA P rapport C 8268-11 (64), Stockholm: Försvarets forskningsanstalt (FOA), Planeringsbyrån, Stockholm, 1971.

Pascoe, Maria, och David Ullner. ”VOICEover, Att automatisk aktivera en passiv sats i svenskan.” Kandidatuppsats i datalingvistik, Institutionen för Lingvistik, Datalingvistikprogrammet, Göteborgs universitet, Göteborg, 2006.

Platzack, Christer. *Svenskans inre grammatik - Det minimalistiska programmet. En introduktion till modern generativ grammatik.* Lund: Studentlitteratur, 1998.

Pollard, Carl, och Ivan A. Sag. *Head-Driven Phrase Structure Grammar.* Chicago: University of Chicago Press, 1994.

”PowerSet.” http://www.powerset.com/ (använd den 01 12 2009).

Ramshaw, Lance, och Mitchell Marcus. ”Text chunking using transformation-based learning.” *Proceedings of Third Workshop on Very Large Corpora.* Somerset, New Jersey: Association for Computational Linguistics, 1995. 82–94.

Ranta, Aarne. *Type-Theoretical Grammar.* Oxford University Press, 1994.

Rosell, Magnus. ”Automatisk synonymvariering av text.” Kursrapport, Språkgranskningsverktyg, KTH, Stockholm, 2005.

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

Rue, Henrik. ”Danish field grammar in typed prolog.” *Proceedings of the third conference on European chapter of the Association for Computational Linguistics.* Köpenhamn: Association for Computational Linguistics, Morristown, NJ, 1987. 167 - 172.

Sag, Ivan A, och Thomas Wasow. ”Syntactic Theory - A Formal Introduction, Partial Draft of September 1997.” Ivan A Sag & Thomas A Wasow, CSLI Publications, Leland Stanford Junior University, 1997.

Sandqvist, Carin. *Studier över meningsbyggnaden i färöiskt skriftspråk (Doktorsavhandling).* Lund: Walter Ekstrand Bokförlag, 1980.

Sigurd, Bengt, och Barbara Gawrońska. *Nexus grammar.* Working papers 42, s 209-224, Lunds universitet, 1994.

Sjöström, Sören. ”Prepositionen i svenskan - öppen eller sluten ordklass?” *Svenskans beskrivning 15.* Göteborg: Sture Allén, Lars-Gunnar Andersson, Jonas Löfström, Kerstin Nordenstam och Bo Ralph, 1985. 474-488.

Sköld, Tryggve. ”Avgör ordföljden vad som är subjekt i en svensk sats?” *Förhandlingar vid sammankomst för att dryfta frågor rörande svenskans beskrivning 3, 1965.* Göteborg: Sture Allén, Inst för nordiska språk, Inst för engelska språket, Göteborgs universitet, 1966. 103-110.

Sneiders, Eriks. *Automated Question Answering: Template-Based Approach (Doktorsavhandling).* Stockholm: Stockholms universitet/KTH, 2002.

Sofkova Hashemi, Sylvana. *Automatic Detection of Grammar Errors in Primary School Children's Texts (Doktorsavhandling).* Göteborgs universitet: Institutionen för lingvistik, 2003.

Spetz, Einar. ”http://www.regionbiblioteket.se/upload/_Dokument/wikipedia.pdf.” *Wikipedias fel och brister - en angelägenhet för biblioteken?* 2008. http://www.regionbiblioteket.se/upload/_Dokument/wikipedia.pdf (använd den 12 10 2009).

”Språkrådets hemsida.” http://www.spraknamnden.se/sprakladan/help%5Ckategorier.htm#annanstrykning (använd den 01 12 2009).

Stroh-Wollin, Ulla. ”Det är något som inte stämmer - om analysen av utbrytningar.” *Svenskans beskrivning 25, 2001.* Åbo: Marketta Sundman och Anne-Marie Londen, 2002. 272-283.

—. *Koncentererad nusvensk formlära och syntax - Övningar med facit.* Lund: Studentlitteratur, 1998.

”Svenska Wikipedia.” http://sv.wikipedia.org (använd den 01 12 2009).

Sågvall Hein, Anna. ”Parsing by means of Uppsala Chart Processor (UCP).” i *Natural Language Parsing Systems*, av L Bolc. Berlin & Heidelberg: Springer Verlag, 1987.

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

—. ”Syntagmer utan namn.” *Förhandlingar vid sammankomst för att dryfta frågor rörande svenskans beskrivning 3, 1965.* Göteborg: Sture Allén, Inst för nordiska språk, Inst för engelska språket, Göteborgs universitet, 1966. 111-116.

—. ”Var går gränsen mellan huvudverb och hjälpverb?” *Svenskans beskrivning 20, Umeå December 1993.* Lund University Press, Studentlitteratur, 1994. 360-369.

Teleman, Ulf, Staffan Hellberg, och Erik Andersson. *Svenska akademiens grammatik.* Stockholm: Norstedts ordbok, 1999.

Thorell, Olof. *Svensk grammatik.* Stockholm: Esselte Studium, 1973.

Welin, Carl Wilhelm. ”En parser för parallell syntaktisk och semantisk analys.” Papers from the Institute of Linguistics, Stockholms Universitet, PILUS 31, Stockholm, 1976.

—. ”Nominalfraser i sammanhängande text.” *Förhandlingar vid sammankomst för att dryfta frågor rörande svenskans beskrivning 8.* Lund: Christer Platzack, 1974.

Westman, Margareta. *Bruksprosa - En funktionell stilanalys med kvantitativ metod.* Lund: Liber Läromedel, Gleerups, 1974.

Viberg, Åke. ”Svenskans lexikala profil.” *Svenskans beskrivning 17, 1989.* Åbo: Åbo akademis förlag, 1990. 391-408.

Viberg, Åke, Kerstin Lindmark, Ann Lindvall, och Ingmarie Mellenius. ”The Swedish WordNet Project.” *Proceedings of Euralex 2002.* Köpenhamn, 2002. 407-412.

Wilhelmsson, Kenneth. ”Automatic Variation of Swedish Text by Syntactic Fronting.” *Workshop on NLP for Reading and Writing - Resources, Algorithms and Tools in conjunction with the SLTC Conference.* Stockholm: Webb-publikation: http://spraakbanken.gu.se/personal/sofie/SLTC_2008/, 2008.

—. ”Heuristic Schema Parsing of Swedish Text.” *SLTC 2008.* Stockholm, 2008.

Wirén, Mats. *Studies in Incremental Natural-Language Analysis (Doktorsavhandling)*. Linköpings universitet: Institutionen för datavetenskap, 1992.

Viterbi, Andrew James. ”Error bounds for convolutional codes and an asymptotically optimum decoding algorithm.” *IEEE Transactions on Information Theory 13 (2)*, 1967: 260-269.

Voutilainen, Atro. ”Parsing Swedish.” *Proc. 13th Nordic Conference on Computational Linguistics (Nodalida-01)*. Uppsala, 2001.

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

Index

Se även *Avhandlingens vokabulär och konventioner* först i avhandlingen

Appendix

Viktiga satsled från Mamban och deras motsvarighet

Satsleden, som här är de som finns på huvudsatsnivå, enligt Mamban (Teleman 1974), närmast hämtade från Landqvist (2000) s. 420, är som nämnts en mer detaljerad kategoriuppsättning än detta arbetes Frågan är dock som nämnts om det mindre detaljerade formatet här skulle kunna nå nästan samma detaljnivå enbart med listningar och andra enklare medel. Detta arbete har egentligen *inte* utgått från Mambans etikettering och segmentering, men det har i slutet blivit tydligt att det finns relativt tydliga motsvarigheter.

Etikett i Mamban (Teleman 1974)	Nuvarande motsvarande etikett i kodningen
FV Finit verb	`pfv`
IV Infinit verb	`piv`
SS Subjekt	`subjekt`
FS Formellt subjekt	
ES Egentligt subjekt	`objekt`
OO Direkt objekt	
IO Indirekt objekt (ej PP)	
SP Subjektiv predikatsfyllnad	
OP Objektiv predikatsfyllnad	
VS Subjekt med infinitiv	`subjekt + objekt`
PL Partikel	`pl`
OA Objektadverbial	`adverbial`
RA Rumsadverbial	
TA Tidsadverbial	
KA Komparationsadverbial	
MA Talarattitydadverbial	
AG Agentadverbial	
+A Konjunktionella adverbial	
CA Framhävande adverbial	
NA Negationsadverbial	
AA Övriga adverbial	
FP Fri subjektiv predikatsfyllnad	
VA Varslande adverbial	Del av led eller adverbial
+F Koordinationsfras på satsnivå	Analys i andra led
XX Obestämbar satsdel	Varierande – men ej obestämd

Tabell 56 Ett försök till mappning av de olika kategorierna i Mamban visar hur framförallt adverbialen är uppdelade i många typer. Det verkar som om gränslinjen mellan begränsade och obegränsade adverbial skär mitt i flera av adverbialen, precis som i kategorin *satsadverbial*.

Kodexempel

```
det_finns_en_strandad_preposition = false;

    if(s.EFTERDEL.STRUKTUR_NIVÅ_3!=undefined)
    {
      for(i=0; i < s.EFTERDEL.STRUKTUR_NIVÅ_3.length; i++)
      {
        if(ainb("PP", meningen.POSARR[ s.EFTERDEL.STRUKTUR_NIVÅ_3[i].SLUT ] ) )
        {
            det_finns_en_strandad_preposition = true;
        }
      }
    }
```

Kodexempel 9 Undersökning av huruvida en strandad preposition förekommer i efterdelen (första delen av identifikation av möjlig rektionsframflyttning) av en sats (`s`) görs genom en genomlöpning av segmentstrukturen. Som visas är det 'strukturnivå 3', dvs. resultatet av eventuell sammanfogning i tre steg som undersöks. I fall ett sådant segment har ett sista element som är en preposition (vilket undersöks med hjälp av index för slutet av segmentet och funktionen `ainb(a,b)` som avgör ifall en sträng är en delsträng av en annan) så kan denna inledningsvis antas vara en strandad preposition. Objekten `mening` och s (sats) från `mr` används som här tillsammans ofta.

Tentativa frågeordsmappningar för prepositioner och bisatsinledande led som använts

De följande mycket tentativa uppställningarna av hv-ordmotsvarigheter användes för tester av frågegenerering i avsnitt 5.2. Tolkningen av de motsvarande hv-orden klargörs i avsnitt 5.2. Uppställningarna ska inte ses som slutgiltiga.

Etiketterna som *V1-fråga, AGENT, frågetecken eller "–"* betyder att ingen fråga genereras av programmet när sådant ord inleder ett primärt satsled.

Prepositionsuppmärkta ord i SUC

af	AGENT av vad / vem / vilka	neråt	Vart
alltefter	när / var	nära	var / hur
alltifrån	när / var	nästintill	var / hur mkt
alltsedan	sedan när	oaktat	Pp
ang	I fråga om vad	oavsett	Pp
angående	I fråga om vad	om	Pp
apropos	I fråga om vad	omkring	var mm
apropå	I fråga om vad	ovan	Var
av	AGENT av vad / vem / vilka	ovanför	Var
bak	var	ovanpå	Var
bakom	var	per	Pp
bakpå	var (på vad)	pr	Pp
beträffande	I fråga om vad	på	var / när
bland	var ("bl a" -> inget)	runt	var / när
bortifrån	varifrån	runtom	Var
bortom	var	runtomkring	Var
bortåt	var / hur mycket / hur många...	rörande	Pp

bredvid	var	sedan	När
därinunder	var	sen	När
efter	när (var)	til	vart / när
emellan	var / när / hur	till	när / vart
emot	varåt / hur	tills	när / vart
enligt	enligt vem / vad / vilka	trots	Pp
framemot	när / var	tvärsöver	Var
framför	var ("framför allt" -> inget)	tvärtemot	Pp
framåt	när	typ	Pp
framöver	när	tä' (till)	vart / när
från	från vad / från när	undan	Var
frånsett	frånsett vem / vad / vilka	under	var / när
för	för vem / vad / vilka	uppemot	hur mkt/många
förbi	var / förbi vad	uppför	Vart
före	när / (var) / före vad	uppifrån	varifrån (pp) / hur
förrän	när	uppåt	Var
förutan	förutan vad	ur	Varifrån
förutom	förutom vad	utan	Pp
genom	hur / var	utanför	Var
gentemot	var / hur	utanpå	Var
hos	var	utav	Pp
i	var / när	utefter	Var
ibland	?	utför	Vart
ifrån	varifrån / närifrån (är piedp-varianter...)	uti	Var
igenom	vart	utifrån	hur / var
inemot	vart / hur mkt	utmed	Var
inför	när	utom	Pp
inifrån	varifrån (är piedp-varianter...)	utur	Vart
inkl.	hur/pp	utåt	Var
inklusive	hur/pp	utöver	Pp
innan	när / var	via	Hur
innanför	när / var	vid	Var
inom	var	wid	Var
inpå	hur / när	à	Pp
intill	var	å	pp (var)
inunder	var	åt'	?
inuti	var	åt	Vart
invid	var	öfver	Var
inåt	var	öfwer	Var
jämte	var	öm	?
kontra	pp	över	VAR
kring	var / hur (mkt)		
liksom	pp		
me	hur		
med	hur		
medelst	hur		
medio	?		
mellan	var / när		
mittemot	var		
mot	var		
mällan	var		
mö (med)	hur		
nedanför	var		
nedför	var / vart		
nedifrån	Var		
nedåt	Vart		
nerför	var varifrån (= pp)		

nerifrån	var varifrån (=)		

Subjunktionsmärkta ord i SUC

allteftersom	när / hur
alltmedan	när / hur
alltsedan	sedan när / hur
at	vad
att	vad (varför)
bara	[V1-fråga eller inget]
eftersom	varför
ehuru (även om)	[V1-fråga eller inget]
emedan (medan)	[V1-fråga eller inget]
enär (eftersom)	varför
fast	[V1-fråga eller inget]
fastän	[V1-fråga eller inget]
för	? Inget
förrän	när
huruvida	vad
ifall	vad / V1-fråga / "under vilka förutsättningar"
innan	när
liksom	[V1-fråga eller inget]
medan	när / V1
mens	när / V1
närhelst	när
om	vad / V1 / "under vilka förutsättningar"
sedan	när
sen	när
så	varför
såvida	[V1-fråga eller inget]
såvitt	[V1-fråga eller inget]
tills	tills/när/tills när
utifall	[V1-fråga eller inget]
än	Inget

HS-märkta ord i SUC

hvars	-
vars	-
vems	vad
vilkas	vad

HA-märkta ord i SUC

der	-
dit	vart
där	var
därför	?därför som?
då	varför/när
heröfver	?
hur	vad
hurdan	vad
hurdant	vad
huru	vad
hurudan	vad
hvad	vad
hvarför	vad
hvarigenom	vad
hwar	vad

hwarifrån	vad
när	vad/när
som	[V1-fråga eller inget]
såsom	[V1-fråga eller inget]
va	vad
vad	vad
vaffer	vad(/adv)
var	vad
varav	-
varefter	[V1-fråga eller inget]
varför	vad / (adv)
varhelst	var
vari	-
varifrån	vad
varigenom	vad
varmed	vad / (med vad)
varpå	-
varstans	-
vart	vad
vartefter	[V1-fråga eller inget]
vartill	-
varur	-
varvid	-

HD-märkta ord i SUC

hurdana	vad
vilka	vad
vilken	vad
vilket	vad (jfr t.ex. "vilket gladde många" - appositionell relativ bisats)

HP-märkta ord i SUC

hvem	vad
hvilken	vad
hwilken	vad
vem	vad
vilken	vad
hvilka	vad
hwilka	vad
vilka	vad
hkt (vilket)	-
hvad	vad
hvilket	vad
hwad	vad
va	vad
vad	vad
vilket	vad
Som	-

Finita anföringsverb som använts

I flera av de ordlistningar som följer ingår ord som visar hur ordgruppen är en nybildande eller i alla fall svårtäckt ordkategori. Anledningen till att sådana ord, t.ex. *tanumspolitikern*, ändå listas som en persontitel är att insamlingen kan ligga till grund för allmänna regler som använder suffixmetodik – så är t.ex. fallet med persontitlar som även känns igen genom suffix. Somliga av listningarna

innehåller olika fel som hamnat där av misstag, eller som placerats där på mindre allmängiltiga grunder. Syftet med uppvisandet här är att visa under vilka förutsättningar som frekvensuppskattningar och resultat tagits fram.

Den första uppställningen innehåller finita anföringsverb som användes under körningar och tester. Det finns här många fall av preteritum- eller presensförekomst där den andra böjningsformen saknas. Observera att felen i nedanstående listningar visar de faktiska förutsättningarna.

anförtrodde, anförtror, anser, anses, ansåg, ansågs, antar, antas, antecknade, antecknar, antog, antogs, argumenterade, argumenterar, avbryter, avbröt, avfärdade, avfärdar, aviserade, aviserar, avrundade, avrundar, avslutade, avslutar, avslöjade, avslöjar, avvärjde, avvärjer, bad, basunerade, basunerar, bedyrade, bedyrar, bedömde, bedömer, befallde, befaller, befarade, befarar, begrep, begriper, beklagade, beklagar, bekräftade, bekräftar, bekände, bekändes, bekänner, bekänns, bemöter, bemötte, beordrade, beordrar, ber, , berättade, berättades, berättar, berättas, berömde, berömmer, beskrev, beskrevs, beskriver, beskrivs, beställde, bestämde, bestämdes, bestämmer, besvarade, besvarar, betonade, betonar, betydde, betyder, bokstaverade, bokstaverar, brummade, brummar, bräker, bräkte, brölade, brölar, bröt, bubblade, bullrade, bullrar, bönar, började, börjar, citerade, citerar, definierade, definierar, deklamerar, deklarerade, deklarerar, dikterade, dikterar, diskuterade, diskuterar, domderade, domderades, domderar, domderas, drev, dundrade, dundrar, envisades, envisas, erbjöd, erfar, erfor, erkände, erkändes, erkänner, erkänns, exemplifierar, fantiserade, fantiserar, fastslog, fastslår, fattade, fattar, filosoferade, filosoferar, flinade, flinar, flämtade, flämtar, flåsar, flög, fnissade, fnissar, fnittrade, fnittrar, fnyser, fnyste, fnös, fortsatte, fortsätter, framhärdade, framhärdar, framhåller, framhöll, framkastade, framställde, framställer, fruktade, fruktar, frustade, frustar, frågade, frågar, funderade, funderar, förbjuder, förbjöd, förebrådde, förebrår, föreslog, föreslår, förespår, förklarade, förklarades, förklarar, förkunnade, förkunnar, förmanade, förmanar, förmodade, förmodar, förmodas, förnekade, förnekar, förstod, förstår, försvarade, försvarar, försäkrade, försäkrades, försäkrar, försäkras, försöker, försökte, förtydligade, förtydligar, förutspådde, förutspår, gallskrek, gallskriker, garvade, garvar, genmälde, , genskjuter, gensköt, gentar, gentog, gissade, gissades, gissar, gissas, gnydde, gnyr, gnällde, gnäller, gormade, gormar, gratulerade, gratulerar, grinade, grinar, grumsade, grumsar, grymtade, grymtades, grymtar, grymtas, grät, gråter, gäspade, gäspar, heter, hetsade, hetsar, hette, hickade, hickar, hojtade, hojtar, hoppades, hoppas, hotade, hotar, hulkade, hulkar, hyssjade, hyssjar, hälsade, hälsar, härmade, härmar, hävdade, hävdades, hävdar, hävdas, hånade, hånar, hör, hörde, hördes, imiterade, imiterar, inflikade, inflikar, informerade, informerades, informerar, informeras, inföll, inledde, inleder, inpräntade, inpräntades, inpräntar, inser, insisterade, insisterar, inskjuter, insköt, instruerade, instruerar, instämde, instämmer, insåg, intalade, intalar, intygade, intygar, invigdes, invände, invänder, ironiserade, ironiserar, jublade, jublar, jämrade, jämrar, kalkylerade, klagade, klagar, klargjorde, klargjordes, klargör, klargörs, kluckade, klämde, klämmer, knarrade, knarrar, knockade, knorrade, kommenterade, kommenterar, konkluderar, konkluderade, konstaterade, konstaterades, konstaterar, konstateras, kontrade, kontrar, krävde, kräver, kved, kvider, ler, ljuder, ljuger, ljöd, ljög, log, lovade, lovar, lugnade, lugnar, lutade, lyder, läser, läspade, läspar, läste, låtsades, låtsas, löd, malde, manade, manar, meddelade, meddelar, medgav, medger, medlade, menade, menades, menar, mimade, mindes, minns, misstänker, misstänkte, morrade, morrar, motiverade, motiverar, mumla, mumlade, mumlades, mumlar, mumlas, murvlade, muttrade, muttrar, myser, myste, märker, märkte, mässade, mässar, nekade, nekar, nickade, nickar, njöt, noterade, noterar, nynnade, nynnar, nämnde, nämner, omtalar, ordlekte, pep, piper, poängterade, poängterades, poängterar, poängteras, pratar, preciserade, predikade, predikar, presenterar, presenteras, prisade, prisar, protesterade, protesterar, pustade, pustar, påminde, påmindes, påminner, påpekade, påpekades, påpekar, påpekas, påstod, påstår, påstås, påtalade, påtalades, påtalar, påtalas, rabblade, rabblar, raljerade, raljerar, rapporterade, rapporterar, redogjorde, redogör, redovisade, redovisar, reflekterade, reflekterar, rekommenderade, rekommenderar, replikerade, replikerar, representerar, resonerade, resonerar, retades, retas, ropade, ropades, ropar, ropas, ryter, rättade, rättar, rådde, råder, råkade, röt, Sa, sade, sades, sammanfattade, sammanfattar, sas, siade, siar, sjunger, sjöng, skojade, skojar, skrattade, skrattar, skrek, skreks, skrev, skrevs, skriker, skriks, skriva, skriver, skrivs, skrockade, skrockar, skryter, skröt, skämtade, skämtades, skämtar, sköt, slutade, småler, smålog, snyftade, snyftar, snäser, snäste, spekulerade, spekulerar, sporde, spådde, spår, stammade, stammar, stortjuter, stämde, stönade, stönar, suckade, suckar, summerade, summerar, svara, svarade, svarades, svarar, svor, svär, säger, säjer, tackade, tackar, talade, talar, teaterviskade, teg, tilla, tillade, tillfogade, tillfogar, tillägger, tillstod, tillstår, tillägger, tjatar, tjöt, tolkar, torde, trevade, trodde, tror, trugade, tröstade, tröstar, tvivlade, tycker, tycks, tyckte, tystade, tystnade, tänker, tänks, tänkte, tänktes, understryker, underströk, undersöker, anförtrodde, anförtror, anser, anses, ansåg, ansågs, antar, antas, antecknade, antecknar, antog, antogs, argumen-

terade, argumenterar, avbryter, avbröt, avfärdade, avfärdar, aviserade, aviserar, avrundade, avrundar, avslutade, avslutar, avslöjade, avslöjar, avvärjde, avvärjer, bad, basunerade, basunerar, bedyrade, bedyrar, bedömde, bedömer, befallde, befaller, befarade, befarar, begrep, begriper, beklagade, beklagar, bekräftade, bekräftar, bekände, bekändes, bekänner, bekänns, bemöter, bemötte, beordrade, beordrar, ber, berättade, berättades, berättar, berättas, berömde, berömmer, beskrev, beskrevs, beskriver, beskrivs, beställde, bestämde, bestämdes, bestämmer, besvarade, besvarar, betonade, betonar, betydde, betyder, bokstaverade, bokstaverar, brummade, brummar, bräker, bräkte, brölade, brölar, bröt, bubblade, bullrade, bullrar, bönar, började, börjar, citerade, citerar, definierade, definierar, deklamerar, deklarerade, deklarerar, dikterade, dikterar, diskuterade, diskuterar, domderade, domderades, domderar, domderas, drev, dundrade, dundrar, envisades, envisas, erbjöd, erfar, erfor, erkände, erkändes, erkänner, erkänns, exemplifierar, fantiserade, fantiserar, fastslog, fastslår, fattade, fattar, filosoferade, filosoferar, flinade, flinar, flämtade, flämtar, flåsar, flög, fnissade, fnissar, fnittrade, fnittrar, fnyser, fnyste, fnös, fortsatte, fortsätter, framhärdade, framhärdar, framhåller, framhöll, framkastade, framställde, framställer, fruktade, fruktar, frustade, frustar, frågade, frågar, funderade, funderar, förbjuder, förbjöd, förebrådde, förebrår, föreslog, föreslår, förespår, förklarade, förklarades, förklarar, förkunnade, förkunnar, förmanade, förmanar, förmodade, förmodar, förmodas, förnekade, förnekar, förstod, förstår, försvarade, försvarar, försäkrade, försäkrades, försäkrar, försäkras, försöker, försökte, förtydligade, förtydligar, förutspådde, förutspår, gallskrek, gallskriker, garvade, garvar, genmälde, genskjuter, gensköt, gentar, gentog, gissade, gissades, gissar, gissas, gnydde, gnyr, gnällde, gnäller, gormade, gormar, gratulerade, gratulerar, grinade, grinar, grumsade, grumsar, grymtade, grymtades, grymtar, grymtas, grät, gråter, gäspade, gäspar, heter, hetsade, hetsar, hette, hickade, hickar, hojtade, hojtar, hoppades, hoppas, hotade, hotar, hulkade, hulkar, hyssjade, hyssjar, hälsade, hälsar, härmade, härmar, hävdade, hävdades, hävdar, hävdas, hånade, hånar, hör, hörde, hördes, imiterade, imiterar, inflikade, inflikar, informerade, informerades, informerar, infor, undervisade, undervisar, undrade, undrar, uppgav, uppger, upplyser, upplyste, uppmanade, uppmanar, upprepade, upprepar, utbrast, utbrister, utlovar, utläser, utläste, utropade, utropades, utropar, uttrycker, uttryckte, utvecklade, utvecklar, varnade, varnar, vidhåller, vidhöll, viskade, viskades, viskar, viskas, visste, vrålade, vrålar, vädjade, vädjar, vägrade, väser, väste, ylade, ylar, återtog, återupptog, önskade, önskar, översatte, översätter

Finita hjälpverbsliknande verb som använts

Den nedanstående uppställningen innehåller de ord som räknats som finita hjälpverb enligt den utvidgade betydelsen och som faktiskt använts.

anges, anses, ansågs, antas, avser, avses, avsåg, bedöms, befanns, befaras, befinnas, behagade, behagar, behövde, behöver, bekräftar, beräknas, beslutade, beslutar, beslutat, beslöt, borde, bredvid, brukade, brukar, bör, började, börjar, börjat, erbjöds, fick, fortsatte, fortsätta, fortsätter, får, föranlett, förbinder, föredrar, föredrog, förefaller, föreföll, föreslås, förmodas, förmådde, förmår, förstod, förstår, försök, försöker, försökt, försökte, förtjänade, förtjänar, förutsattes, förutsätts, förväntades, gick, gillade, gillar, gitte, gitter, glöm, glömde, glömmer, gällde, gäller, går, hade, hann, har, hatade, hinner, hjälptes, hoppades, hoppas, hotade, hotar, hunnit, håller, ingick, kan, klarade, klarar, kom, kommer, kräver, kunde, lovade, lyckades, lyckas, lyckats, lär, låtsades, låtsas, misslyckades, misslyckas, mäktade, mäktar, må, månde, måste, måtte, nännas, nänns, nödgades, nödgas, orkade, orkar, planerade, planerar, plägade, plägar, prova, provade, provar, prövade, prövar, påstås, rekommenderas, riskerade, riskerar, räknas, råkade, råkar, sades, sas, ses, ska, skall, skulle, slapp, slipper, slutade, slutar, svär, synes, syns, syntes, sägs, sågs, söker, tenderade, tenderar, testade, testar, tillåts, torde, tordes, tros, turades, tvangs, tvekar, tvingade, tvingades, tvingar, tvingas, tycks, tycktes, tänker, tänkte, tål, tålde, törs, undgick, undgå, undgår, undvek, undviker, uppfattades, uppfattas, upphör, upphörde, uppmanas, valde, verkade, verkar, vill, ville, visade, visste, vägrade, vägrar, väljer, väntades, väntas, värdesatte, vågade, vågar, wille, ämnade, ämnar, åtar, återstår, önskade, önskar, övervägde, överväger

Nomen-ord med potentiell adverbialfunktion som använts

Nedanstående ord, främst substantiv, gäller sådana som använts för att markera att när de förekommer som huvudord i nominalfraser kan låta frasen ha adverbialfunktion. De flesta förekomsterna av sådana adverbial är tidsadverbial.

annandag, annandagen, apr, april, aug, augusti, avtalsåret, barndomstiden, barndomsåren, bit, budgetåret, da, dag, dagar, dagarna, dagen, dagligen, dags, danskväll, dar, dec, december, decennier, decenniet, decennium, dygn, dygnen, dygnet, eftermiddag, eftermiddagar, eftermiddagarna, eftermiddagen, fastan, feb, februari, fjol, fjolåret, fre, freda, fredag, fredag-söndag, fredagen, förkväll, förkvällar, förkvällen, förmiddag, förmiddagar, förmiddagarna, förmiddagen, försommardag, grand, gryning, gryningen, gång, gången, gånger, gångerna, halvlek, halvtimme, halvår, halvåren, halvåret, helg, helger, helgerna, hvardagar, häromåret, höst, höstdag, hösten, höstkvällar, höstterminen, jan, januari, jul, juli, juni, juninatt, kilometer, klockan, km, kvart, kvartal, kvartalen, kvartalet, kvarten, kväll, kvällen, kvällning, kvällningen, lektionstimmar, liv, livet, lunchtid, läsåren, läsåret, lördag, lördagsmorgon, maj, mars, metrarna, midnatt, midsommar, midsommaren, midsommarn, mil, minut, minuten, minuter, minuterna, morgon, morgonen, mornar, morron, morse, mån, månad, månaden, månader, månaderna, måndag, måndagarna, måndagen, natt, natten, nov, nov., november, nyår, nyåret, nätter, nätterna, okt, oktober, ons, onsdag, onsdagen, period, perioden, perioder, perioderna, pingst, påsk, sekund, sekunddelar, sekunder, semestern, senare, sensommardag, sep, sept, september, sextiotalen, skärtorsdagen, slag, sommar, sommardag, sommaren, sommarn, sommartid, somrar, somrarna, stenkast, stund, stunden, stunder, stunderna, stycke, säsong, säsongen, säsonger, säsongerna, sön, söndag, söndagen, söndagskvällen, tag, termin, tid, tiden, tider, tidpunkt, tidpunkten, tidpunkter, tidpunkterna, tillfälle, tillfällen, tillfällena, tillfället, timma, timmar, timmarna, timme, timmen, tis, tisdag, tisdagen, tor, torrår, tors, torsdag, trappa, trappsteg, trappsteg, vargtimma, vargtimme, vecka, veckan, veckor, veckorna, vinter, vinterdag, vinterkväll, vinterkvällar, vintern, vintertid, väg, vår, vårdag, våren, vårhelg, vårkväll, vårkvällar, vårkvällarna, vårkvällen, vårterminen, vårvintern, år, åren, året, årsskiftet, årtiondena, årtusende, årtusenden, årtusendet, ögonblick, ögonblicken, ögonblicket

Persontitlar som använts

Nedanstående listning gäller de ord som urskiljs som attributiva före egennamn. Denna uppdelning är viktig för att kunna sammanfoga två ibland åtskilda segment. I listningen ingår andra heuristiskt motiverade ord som fungerar på samma sätt. Den syntaktiska funktionen för denna ordgrupp är framför allt att de sammanfogas med framförvarande ord när detta är egennamn. Dessutom markerar ordgruppen oftast animathetsvärdet *mänsklig*.

40-åringen, abb, abbé, ackompanjatör, actionstjärnan, adelsmannen, adjunkt, adjunkten, adress, advokat, advokaten, affärsmagasinet, affärsmannen, affärstidningen, agitatorn, airedaleterriern, akrylmålningen, aktiebolaget, albumet, algen, ambulansföraren, amerikaner, amten, analysinstitutet, anfader, anstaltschefen, apso-kenneln, araben, arbete, arbetsförmedlaren, arbetsgivare, arbetsgruppen, arbetskamrater, arbetsledaren, arkansasguvernören, arkitekten, arosduon, arrangörsklubben, artikeln, artistnamnet, arvfienden, arvtagare, assistent, astronomen, auktionsfirman, avdelningsdirektör, avdelningskollega, avelshunden, avelsmatadoren, avhandling, ayatollah, backkämpen, bagaren, bakhjulsupphängningen, bandelen, bandhundar, bankdirektören, bar, barn, barnbarn, barnfigur, baron, baronessan, basen, begreppet, berwaldeleven, betjänt, betjänten, betonglagbas, bildningarna, bilkunderna, biografen, biografi, biskop, biskopen, boken, bolagen, bolaget, bolsjevikledaren, bordtennisspelaren, borgmästare, boställe, botanisten, brandchefen, branschtidningen, broder, brodern, bror, brorson, brottsling, bråkstaken, bröderna, bröderne, bulgariskan, butikschefen, byarna, byhövdingen, byn, börschefen, börsfavoriten, café, cembalisten, centerledaren, chauffören, chef, chefen, chefjurist, chefsekonom, chefsåklagare, chowen, citatet, civilingenjör, clash-gitarristen, commander, count, cyrenaren, daghemmet, dalai, dame, dansarna, dans-

ken, dataföretaget, datajätten, datakonsulten, delstat, delstaten, departementsledarna, departementssekreterare, destination, detektiven, diakonen, dialogen, direktör, dirigent, dirigenten, dissidenten, div, divisionschefen, djursholmsscouten, docent, docenten, doktorand, doktorsavhandling, dokumentet, domare, domaren, don, dotter, dotterbolag, dotterbolaget, dr, dr., dragspelaren, drottning, dubbelmonarkin, döttrar, döttrarna, ecklesiastikminister, effekten, efta-staterna, efterträdare, ekochefen, ekonomichef, ekonomidirektör, elektronpositronkollideraren, emeritus, engelsmannen, enheten, enmansutredaren, ensemble, epitetet, epoken, esten, etnologen, exegeten, fader, fadern, fallet, familjen, far, farbrodern, farbror, farmor, faster, fastigheterna, fastighetsföretaget, fastighetsingenjör, femman, ff-spelaren, fil, filmen, filmforskare, filmstjärnan, filosofen, finansborgarrådet, finansminister, finländaren, finskan, fiskaren, fiskerikonsulent, fixstjärna, fjädringen, flaggregistren, flickan, flickfavoriten, flickvän, floderna, flygplanet, flöjtisten, fn-medlaren, foa-forskaren, folkbildaren, folkfronten, folkpartiet, forskaren, forskningsföretag, forskningsingenjör, forskningsorganisationen, fotbollsnationerna, fotbollsspelare, fotbollsspelaren, fotograf, fotografen, framtidsföretaget, fransman, fransmannen, fransmän, fransyskan, frau, friherre, fru, fruarna, fruktjätten, fröken, fysikern, fysikerna, fängelset, fänrik, fänriken, färdledaren, färtmarskalk, förbrytaren, förbundskansler, förbundskapten, förbundskaptenerna, förbundsrepubliken, föreningen, föreståndaren, företagen, företaget, företagsekonomen, företrädare, förfader, författaren, författarinnan, förkunnaren, förkämpe, förläggare, förnamnet, försvarsminister, försäkringsbolaget, förvaltningschef, fürst, gabonskan, gaistränaren, galleri, galleriägaren, gardisten, gatuchefen, general, generaldirektören, generalmajor, geologen, germanisten, gimochefen, gitarrduon, glasblåsaren, gnomonen, golden, golfexpert, gossen, granatgevär, granatgeväret, grannarna, grannen, grannland, grannlandet, grannorten, grannsocknar, granskningsingenjör, greve, grevinnan, grosshandlaren, grossistförbundet, gruppen, grusspecialisten, grusspecialisterna, gubbe, gudaborgen, guden, gummiverkstadsföretaget, gävleflickan, gården, göteborgarna, göteborgsfirman, göteborgsprofessorn, hamnstaden, haveridirektör, hemlandet, hemmahoppet, hemmamålvakten, hemmanationen, hemmasonen, hemstad, herr, herrar, herrn, hertig, historikern, hjulmakaren, hockeykämpe, hockeyspelaren, hollywood-veteranen, hollywoodstjärnan, hotel, hotell, hovreporter, hovskald, hovsångerskan, hr, humanisten, husgudarna, hustru, hustrun, huvudkontoret, huvudpersoner, huvudsekreteraren, huvudstad, huvudstaden, häraderna, hårdrockarna, högerpartiet, höghastighetståget, ibm-forskarna, ifk, iii-laget, indianhövdingen, indiern, industriministern, ingenjör, inrikeschef, insektsforskaren, instrumentmästaren, intendent, invandrarminister, isbrytaren, israelen, italienaren, iv-laget, japanen, japanskan, jazzguru, jonglören, joniern, jordbrukspolitiker, journalisten, jungfru, junioren, juniortränarna, juristen, järnvägarna, järnvägsföretagen, kadett, kaféet, kammarrättsassessor, kammaråklagare, kampanj, kamrat, kamraten, kamrer, kanslist, kap., kapten, kaptenen, kardinalen, karnevalsgeneralen, karolinska, kassör, katten, kattkvinnan, kds-aren, kds-ledaren, kds:aren, kedjekamrat, kejsar, kejsaren, kejsarinnan, klassföreståndare, klassificeringsbolaget, klassiker, klostret, klubbdirektören, klubbkamrat, klubbkamraten, klubbordförande, knarkkungen, kocken, kollega, kollegan, kombinationsförsäkring, komedin, kommendör, kommendörkapten, kommentatorn, komminister, kommissarie, kommittén, kommunalpolitikerna, kommunalrådet, kommundirektör, kommunerna, kommunikationsminister, kompisarna, kompisen, kompositören, koncernchef, koncernchefen, konditori, konkurrenten, konkurrenterna, konspiratören, konstnären, konstprofessorn, konstvetaren, kontingensteoretikern, kontot, konung, koreografen, korpral, krigsmaterielinspektör, kriminalinspektör, kriminalkommissarie, kriminalkommissarien, kronor-basen, kronprinsessan, kulturföreningen, kultursekreterare, kung, kungen, kungl, kurator, kursgård, kusin, kustfylkena, kvadraten, kvarnen, kvarteret, kvarterspolis, kyrkofadern, kyrkoherde, kyrkoherden, kärnkraftverket, kårordförande, kökarflickan, köpingorten, körlyrikern, lady, lagen, lagkapten, lagverk, lake, lakelandterriern, landet, landshövding, landslagsmannen, landslagsstjärna, landslagstjejen, landsman, landsvägen, lantbrukaren, lappskattelandet, ledamot, ledarduon, ledare, legenden, les, liberalen, lillasyster, lindsay-hoggs, linear, lingua, lingvisten, linjen, liraren, livsverk, ljusdalsbon, ljusdesignern, ljussättare, lo-mannen, lokalavd, lokaltrafik, lord, lotsarna, lundafilosofen, lundahistorikern, lyrikboken, lyxhotellet, läkaren, läkemedlet, länet, lärare, läraren, lärjunge, läromästare, lågprisbutiken, lågstadieläraren, långivare, låten, löjtnant, lösdrivaren, m/s, madame, maestro, magister, majestät, major, maken, mamma, manager, manusförfattaren, marknadschefen, marknadsstaden, maskinfirman, matematikern, mb-ledamoten, medarrangören, medeltidsstaden, medlet, megastjärnan, meningen, meteorologen, metropolen, miljonstaden, miljöombudsman, minnesboken, minnesforskaren, miss, missionären, mister, mittback, mittbacken, moderatledaren, monsieur, mont, monte, mor, morbror, morfar, morgontidningen, moselle, moselles, moster, motparten, motståndare, mount, mr, mr., mrs, ms, munken, murarlärlingen, murarna, musikalen, musikdirektör, musikerna, musikinstrumentet, människan, märkena, märket, mästaren, mäster, mästersmeden, målaren, målvakten, månadstidningen, nacka, namn, namnbeteckningen, nationalisten, nattågsprodukten, naturreservatet, naturväktaren, new, nokia, nordic, notre, nova, novellen, nunnan, nyhetsbyrån, nyhetsmagasinet, nykomlingen, nykterhetsorden, oljetätningsföretaget, ombudsman, området, opera, operan, operation, operativsystemet, oppositionspolitikern, ordförande, ordföranden, ordidisslaren, organisationen, orkanen, orkidén, ormen, orten, os-laget, p1-dokumentär, paddan, pappa, paret, parollen, partiledaren, partisekreterare, partisekreteraren, passat, pastor, pekingesen, pekoralet, pensionär, pensionären, personaldirektör, pianisten, piga, pigroman, pingsthelgen, place, platschefen, poeten, pojken, pojkvasker, pojkvännen, polisassistent, poliskommissarie, polismannen, polismästare, polisuppsyningsman, politikern, poor, port, postdirektören, postfröken,

premiärminister, premiärministern, presentatören, presidenten, presidentkandidaterna, pressattachén, prince, prins, prinsessan, producenten, produktionschefen, produktutvecklarna, prof, prof., professor, professorn, professorskan, profeten, proffshandlern, programledare, programledaren, programledarna, programsättaren, projekt, projektet, prosten, provinsen, provinser, prästen, prästmännen, ps, psykiatern, psykoanalytikern, pyromanen, pythagorén, pyttelandet, påven, père, queen, racingentusiasten, raketbasen, ransarkungen, rapport, rasen, redaktören, regeringen, regeringschefen, regeringsorganet, regionen, regissören, rektor, renault, rennäringskonsulent, renässansfursten, reportageboken, republiken, restaurang, restaurangen, revisorsfederationen, revysuccén, riksbankschef, riksbankschefen, riksdagsledamoten, rikskansler, riksmedlaren, rikspolischef, rikspostmästarinnan, rocksångaren, roman, romanen, romaren, rubriken, rue, rumsförmedlingen, rumänen, runstensområdet, rutten, rymdbasen, rymdfärjan, rymdorganet, rymdorganisationen, rymdstationen, ryssen, ryttmästare, rådsherre, río, salon, samarbetsorganen, samarbetsorganisation, sambon, samhället, samordnare, sandinisttidningen, sankt, sassanidkonungen, scandinavian, scenarbetaren, scenografen, schejk, schwedischen, schweitzer, schweizaren, scoutdissidenten, segeltävlingen, seglarna, sekreteraren, seminarielärare, senate, senatorn, sergeant, seriefiguren, settern, showgruppen, sidokollisionsskyddet, sif-bas, sir, sistemannen, sj-arkitekten, sj-chefen, sjuksköterskan, sjukvårdsekonomen, sjuttifemma, sjö, sjökaptenen, sjön, skald, skalden, skattedirektör, skeppet, skeppsredarfamiljen, skogsekologen, skoldirektör, skolvaktmästare, skomakare, skotten, skribenten, skulpturen, skulptören, skådespelarna, skådespelerskan, slagverkaren, slaktare, slalom-ess, slottet, slutstation, släktet, släkting, släktkrönikan, släktnamnet, smålandstösen, snickaren, socialdemokratiska, socialminister, socialministern, sociologen, solisten, soloalbum, sommarlovsteatern, sommarpalats, son, sonen, sor, southern, sovjetledaren, spanieln, stabbläggaren, staden, star, starke, staten, stationschefen, statsmannen, statsminister, statsrådet, statssekreterare, stena, stiftelsen, stiga, stinsen, stockholmpolisen, stoikern, stollen, storbandet, storfurstendömet, storköksverksamhet, storpudeln, sträckan, sträckorna, strängnäslektorn, strävhårstaxen, studieförbundet, studierektor, studio, städerna, subaru, succéfilm, sudanesen, sultan, sultanatet, supermakten, suzuki, sveakonungen, svensken, svensktoppsartisten, svärson, syn, syskon, syskonparet, syster, systern, säkerhetstjänsten, sällskapet, särlingen, sågverksorterna, sångaren, sångerskan, sökaren, talesman, tall, talman, tant, tanumspolitikern, team, teater, teaterarkitekten, teatergruppen, teatermannen, telefonsystemet, temat, tenorerna, teoretikern, teoretikerna, teststräckorna, tetra, texasmiljardären, textförfattaren, textilkonstnärinnan, the, theatre, this, thrillern, théatre, tiden-redaktören, tidning, tidningen, tidningstecknaren, tidskrift, tillnamnet, timbroskriften, tjecken, tjejgrabbarna, tjänare, to, tonsättaren, tonåringen, total, toyota, trafalgar, trafik, triosonatan, trombonisten, trubadur, trummisen, trädgårdspedagog, tränare, tränaren, tsar, tullingeveteranen, tunnelborrmaskin, turridningsfirman, tusenstreckskonstnären, tv-program, tvåan, typ, tysken, tätningsföretaget, tätorten, tågmästare, ultra, un, ungdomsorganisationen, ungdomspjäsen, unge, ungraren, uppdrag, uppfödaren, upphovsmannen, uppsala-filosofen, uppsaliensaren, upptäcktsresande, usa-ambassadören, utbildningsföretaget, utrikeshandelsminister, utrikesminister, utrikesministern, utsaga, utvecklingsbolag, v65-debutanten, vadstenamunken, vaktmästar, valkyrian, vallentunaföretaget, vallokomotiv, valutasamarbete, vattendragen, vd, veckotidningen, verkmästare, vestalen, vetenskapshistorikern, veteranen, vicepresidentkandidat, vidundret, violinisten, vissångerskan, vsk-arna, vulkanen, vän, väninnan, vännen, vännerna, vänstertidskriften, världsmästaren, världsmästarinnan, värmlänningen, västeråsföretaget, västra, with, wordperfect, youth, zeppelinaren, zero, zoologen, ägare, ägarna, ämnesinstitutionerna, ämnet, änkan, änkorna, ärkebiskop, ärkediakonen, öarna, öknamnet, ön, örebroträanaren, örlogskapten, örlogsstaden, österrikaren, östra, östtysken, överingenjör, överläkare, överrikena, överriket, övers., överskriften, översköterskan, överste, överstelöjtnant, överstinnan, översättning

Mängdord som använts

Nedanstående uppställning gäller de ord som listats för att sammanfogas med efterföljande NP, speciellt substantiv, på liknande grunder som för uppställningen av persontitlar ovan. Som mängdord finns utöver de tydliga medlemmarna i gruppen, en del ord som på heuristiska grunder fått denna markering (uppställningen är inte felfri).

10-pack, 10-tal, 100-talet, 2-pack, 20-tal, 25-tal, 25-talet, 3-pack, 30-tal, 4-pack, 40-tal, 50-tal, 50-talet, 6-pack, 70-tal, aln, ambassadör, amilj, ampere, andel, andelen, aning, aningen, antal, antalet, appendix, arealen, armada, armé, arsenal, art., artikel, ask, avd, avdelning, avdelningen, avsnitt, avsnitten, avsnittet, back, balja, begrepp, begreppen, begreppet, begreppsparet, benämningen, besättning, beteckningen, bilaga, bilagan, bild, biljon, billi-

on, bit, blad, blandningen, blodfaktor, blodgrupp, bok, brandel, bråkdelen, bukett, bunke, bunt, burk, buss, busslast, butiken, byte, bägare, cal, candela, celsius, centigram, centiliter, centimeter, cl, cm, cm2, cm3, damas, deciliter, decimeter, del, delar, delaren, delkategorierna, delmängd, delresultatet, derrida, dessertsked, diagnos, div, div., division, divisioner, dks, dl, dm, dm2, dm3, doktor, dos, dotterbolag, dottern, droppar, droppe, drottning, dsk, dussin, enh, enhet, enheten, epitetet, fader, faktor, famn, fas, fat, fempack, femtio-tal, femtiotal, femtital, femtontal, fig, figur, figurerna, flaska, flaskan, flaskor, flaskorna, flertal, flertalet, flickvän, flock, flockar, floden, flotta, folkfronten, fot, fpg, fyrpack, fyrtiotalet, fyrtotal, fält, fältet, fång, fåtal, förhållandet, förordning, förpackning, förråd, g, generation, generationen, generationer, genren, glas, gnutta, gram, grenarna, grupp, gruppen, grupper, gäng, ha, hallwylska, halt, halv, halvdussin, halvmeter, handfull, handfulla, handlingen, hektar, hela, hord, hundratal, hundratalet, härskara, härva, ifk, inkomstslaget, iso, joh, kanna, kap, kap., kapitel, kapsel, kartong, kartonger, kasse, kastrull, kategori, kg, kilo, kilobyte, kilogram, kilohertz, kilojoule, kilokalori, kilometer, kilovolt, kilowatt, kilowattimme, kj, klass, klick, klockan, klunk, klyfta, km, km2, knippe, knippen, knippet, kolonn, kolumnen, kombinationen, kompani, konsumentföreningen, kopp, koppar, koppel, korn, krm, kryddmått, kubikcentimeter, kubikdecimeter, kubikmeter, kubikmillimeter, kvadratcentimeter, kvadratdecimeter, kvadraten, kvadratkilometer, kvadratmeter, kvadratmillimeter, kvittningsrätten, kwh, köpmännens, laddning, lag, lager, lass, last, lastbilslass, lastmåttet, leifler, limpa, limpor, linje, linjer, liter, lo-mannen, läsåret, låda, lådan, lådor, lådorna, m/s, m2, m3, magnetiseringskoden, majorsbostället, mark, massa, matsked, matt, mb, mbar, megabyte, megahertz, megajoule, megawatt, megawattimme, merpart, meter, mg, mhz, mikrogram, mikrometer, mil, milj., miljard, miljarder, miljon, miljoner, miljontal, milliard, millibar, milligram, milliliter, millimeter, minut, mixen, ml, mm, mm2, mm3, modell, mom., moment, motsatsparet, msk, mugg, myckenhet, mängd, mängden, mängder, målen, mångfald, mått, n:o, namnen, namnet, nanometer, newton, nikomakos, nr, nummer, nykomlingen, nypa, nypor, nävar, omg, omgång, omgångar, område, opinionsinstitutet, oppositionspartiet, ord, ordet, ordning, ordningen, organisationerna, otal, ounce, p., pack, packe, paket, par, paragraf, paret, parken, parollen, parti, pass, pikometer, pluton, plätt, pojk-, porsche, portion, postnummer, pound, president, principen, proc, procent, procentandel, procentenhet, procentenheter, promille, prop, prop., prästgården, psalm, punkt, punkten, påse, rad, ring, rote, rubrikerna, rulle, ruwi, räcka, s., sal, sambandet, samling, samlingen, sats, schejk, serie, serien, sexa, sexpack, sid, sida, sidan, siffra, siffran, sjuttifemma, sjuttiotal, skala, skara, skaran, sked, skiva, skivan, skivor, skivorna, skock, skoda, skopa, skvadron, skäppa, skål, skålpund, slag, släktet, smula, smulan, smulor, smulorna, sonen, sorten, southern, spann, st, stab, statsskicket, stone, strimma, strolz, strut, studio, styck, stycke, stycken, stycket, ställningen, succéromanen, summa, summan, system, säck, säsongen, tabell, tablett, talet, tallrik, tant, technical, tel, temat, terabyte, terawattimme, termen, terminal, tesked, the, tidskriften, tiopack, tiotal, tiotalet, titeln, tjog, tjogtal, tjugotal, tolv, ton, trave, trepack, trettiotal, trettital, tropp, trupp, tråg, tsk, tum, tunna, tusental, tuva, tvåpack, typ, typen, typer, ugn, ungdomsorganisationen, uppsjö, uppsättning, utrikeshandelsminister, vinkeln, voltampere, volym, waern, wiberg, xxxx000, yard, zon, §, °, åldersgrupp, ångström, år, åren, året, årsklass, årskurs