Institutionen för lingvistik
Stockholms universitet
VT-98

FELAKTIGT SÄRSKRIVNA SAMMANSÄTTNINGAR

Lena Öhrman

I denna uppsats undersöks en samling felaktigt särskrivna sammansättningar som hämtats från Internet. Utifrån dessa skapas en typologi baserad på förledens och efterledens ordklass. Typen *substantiv + substantiv* representerar hela 70% av alla funna särskrivningar. Med hjälp av regler i språkgranskningsprogrammet Granska undersöks möjligheten att automatiskt detektera denna typ av särskrivna sammansättningar. Körning av reglerna tyder på bra resultat för sådana särskrivningar där förledet är ett substantiv och efterledet är ett substantiv i bestämd form.

Påbyggnadskurs i datorlingvistik
Uppsats 10 poäng
Handledare: Gunnel Källgren (institutionen),
Rickard Domeij & Ola Knutsson (IPLab, Nada, KTH)

INNEHÅLLSFÖRTECKNING

1 INLEDNING 3

1.1 SYFTE 3

2 BAKGRUND 3

2.1 SVENSKA SAMMANSÄTTNINGAR 4

2.1.1 Förled och efterled 4

2.1.2 Förledens utseende 5

2.1.3 Ihop eller isär? 6

2.2 SÄRSKRIVNINGAR 7

2.3 SPRÅKGRANSKNINGSPROGRAMMET GRANSKA 7

3 METOD 9

3.1 MATERIAL 9

3.2 TILLVÄGAGÅNGSSÄTT 9

4 RESULTAT 10

4.1 TYPOLOGI ÖVER OLIKA SÄRSKRIVNINGSTYPER 10

4.2 DETEKTION AV FELAKTIGA SÄRSKRIVNINGAR 11

4.2.1 Regler i Granska 12

4.3 SAMMANFATTNING AV RESULTAT 16

5 DISKUSSION 16

5.1 FRAMTIDA ARBETE 17

6 REFERENSER 18

APPENDIX A: SUC-TAGGAR OCH STILTAGGAR 19

APPENDIX B: UTDRAG UR FELSAMLING 21

1 INLEDNING

Det här arbetet görs som en del av ett projekt på Nada vid Tekniska Högskolan i Stockholm där ett program för datorstödd språkgranskning utvecklas. Projektet har som mål att utveckla datorfunktioner för språkgranskning på svenska, att införa dessa i en interaktiv miljö för datorstött skrivande och att testa dem empiriskt med användare. Granskningsfunktionen analyserar språket i en text och markerar de problem den finner i texten.

1.1 Syfte

Syftet med uppsatsen är att undersöka möjligheten att detektera felaktigt särskrivna sammansättningar med hjälp av dator. I uppgiften ingår att analysera och beskriva konstruktionen av olika typer av särskrivningar, ge förslag på lösningar för att automatiskt/interaktivt detektera dem, och, om möjligt, visa att detta går att implementera på ett sätt som är användbart i datorstödd språkgranskning.

2 BAKGRUND

Felaktigt särskrivna sammansättningar är något man ser allt oftare. Engelskan tycks påverka svenska språket på många områden, kanske framför allt i IT-sammanhang. I svenskan, till skillnad från engelskan, utgör felaktig särskrivning av sammansättningar ett språkproblem. Tyvärr är många granskningsprogram amerikanska till sitt ursprung och kan inte hantera denna typ av språkproblem.

Engelskan är nog inte det enda som påverkar till fler särskrivningar. Reklam kan vara en annan orsak. Reklammakarna är förtjusta i så kallad gestalttext som är utformad på ett sätt som gör att vi lägger märke till den: iögonfallande typsnitt, färgglada bokstäver, stor text osv. Under sådana förhållanden är det inte så konstigt om vi lägger märke till textens utformning och påverkas av dess språkbruk. Eftersom den texten ibland skrivs på höjden med avstavningar utan bindestreck, är utfallet givet. I sådan text är syftet inte i första hand läsbarhet utan t.ex. att texten ska väcka uppmärksamhet, se bra ut, förmedla en känsla osv. Då vi utsätts mer och mer för text som gått genom reklammakares händer, påverkas vi undermedvetet och skriver sedan på samma sätt själva (SvD 971207).

Även de som skriver restaurangmenyer verkar ha en viss förkärlek för felaktigt särskrivna sammansättningar. Ibland kan man se dussinet felaktiga särskrivningar på en och samma matsedel. Särskrivningar blir allt vanligare också i löpande text. Man ser dem på Internet, i uppsatser, i e-postmeddelanden och så vidare. Exempel på felaktigt särskrivna sammansättningar: *hus bil, gat lykta, silver tejp*. Knivigare blir det med sammansättningar som får en annan betydelse då de särskrivs som till exempel *rått kött, brun sås, sjuk sköterska*, *rök fritt.* När ska ett ord skrivas ihop respektive isär? Jag har gjort en litteraturundersökning av hur sammansättningar ser ut och bildas i svenska språket som redovisas under avsnittet *Svenska sammansättningar* nedan. Litteraturundersökningen är framför allt baserad på Thorells Svensk Ordbildningslära, Svenska Språknämndens skrivregler och SAOL. Sedan följer en egenhändigt gjord undersökning av felaktigt särskrivna sammansättningar. För att studera särskrivningarna har jag till största delen använt mig av min egen undersökning i vilken jag samlat in 389 särskrivningar med kontexter från Internet.

2.1 Svenska sammansättningar

När ska ett ord skrivas ihop respektive isär? Jag ska börja med att titta på vad en sammansättning är: en sammansättning är ett ord som kan delas upp i minst två (ordliknande) huvuddelar som var och en innehåller minst ett rotmorfem. Ett exempel på den enklaste typen av sammansättningar är de som består av två rotmorfem, som i ordet *jordbruk*.

Typiskt för germanska språk är att de kan bilda nya ord genom sammansättningar, jämför till exempel svenskans *järnväg*, tyskans *Eisenbahn* och engelskans *railway* med franskans *chemin de fer*. I svenskan är möjligheterna att bilda sammansättningar i princip obegränsade. Det finns också sammansättningar som själva ingår i nya, längre sammansättningar, t.ex. *småjordbruk.* Vidare innehåller sammansättningar också fogemorfem. Ett exempel på det är *skrivningsvakt.* (Thorell:32).

För att på ett enkelt sätt åskådliggöra strukturen i en sammansättning kan man rita träddiagram som i figur 1 nedan.

(SMS=sammansättning, AVL=avledning, RM=rotmorfem, FM=fogemorfem, AS=avledningssuffix)

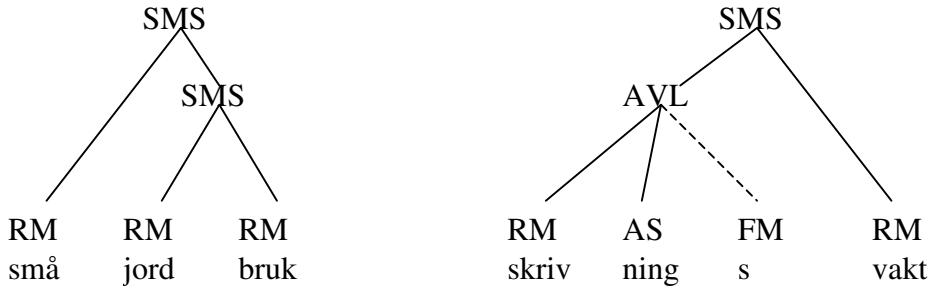

Fig. 1 (Thorell:33)

2.1.1 Förled och efterled

Den första delen av en sammansättning brukar kallas förled och den andra efterled. Mellan dessa förekommer ofta ett fogemorfem. Både förled och efterled är viktiga delar, fast på olika sätt. Efterleden avgör bland annat vilken ordklass sammansättningen tillhör. Om efterledet är ett substantiv (t.ex. *vin*) blir sammansättningen (t.ex. *rödvin*) också ett substantiv. Ofta är sammansättningen hyponym till efterledet (ett rödvin är ett slags vin). Fast ibland stämmer inte detta för en jordgubbe är ju inte en sorts gubbe.

Förleden är i de flesta sammansättningar en sorts bestämning till efterleden, de "determinerar" efterleden.

I fig. 2 nedan visas nio olika kombinationer där de två viktiga delarna, förleden och efterleden, kan vara substantiviska, adjektiviska eller verbala:

EFTERLED:	Subst. (S)	Adj. (A)	Verb (V)
FÖRLED:			
Subst. (S)	barnbok	skogrik	halshugga
Adj. (A)	privatperson	gröngul	storskratta
Verb (V)	läsrum	sittriktig	lästräna

Fig. 2 (Thorell:34)

Förleden kan även bestå av följande ordklasser: pronomen, räkneord och adverb eller preposition. En förlängning av fig. 2 skulle kunna se ut så här:

EFTERLED:	Subst.	Adj.	Verb
FÖRLED:			
Pron.	dusägande	jagbetonad	självdö
Räkneord	tiotal	trestjärnig	tvådela
Adv. el. prep.	baksäte	nygift	avskriva

Fig. 3

Den i särklass vanligaste typen är substantiv + substantiv (SS) och därefter kommer de båda övriga kombinationerna med substantiv som efterled (AS resp. VS) (Thorell:34). Kombinationen verb + verb förekommer sällan.

I vissa fall stöter man på problem med att avgöra om förleden är substantiviska eller verbala. Till exempel *köp* i *köpkraft*. Problemet uppstår eftersom det faktiskt finns ett substantiv *köp*. Däremot borde man kunna säga att *läs* i *lästräna* är ett verb eftersom det inte finns något substantiv *läs*.

Man skiljer på två typer av sammansättningar i svenska, nämligen determinativ och kopulativ. I sammansättningar bestående av adjektiv + substantiv, t.ex. *privatperson*, motsvarar förledet ett adjektivattribut. Dessa sammansättningar kallas determinativa därför att förleden är en bestämning till (“determinerar”) efterleden. Determinativa är den vanligaste typen, mindre vanliga är de så kallade kopulativa. Ett exempel på kopulativ sammansättning är *blågul*, där förledet inte bestämmer efterledet utan snarare är jämställt med efterledet.

2.1.2 Förledens utseende

Förleden kan se väldigt olika ut från fall till fall. Det finns inga heltäckande regler för hur förleden ska se ut men det finns några fall som är vanligare än andra och som jag tänker beröra här.

När förleden är adjektiviska händer oftast ingenting (t.ex. *rödvin, sjukpension*), förleden är lika med motsvarande adjektiv i grundform.

När förleden är verbala är de nästan alltid lika med stammen av motsvarande verb (läsa + rum blir läsrum). Ibland måste ett *-e* skjutas in av uttalsskäl (klättra + ställning blir klätterställning).

Substantiviska förled är väldigt komplexa. Man kan dela upp substantiven i "enkla" substantiv och substantiv som är sammansättningar (Thorell:34). T.ex. är det svårt att veta om enkla substantiv som slutar på konsonant kräver fogeelementet *-s* eller inte. (I äldre sammansättningar förekommer andra fogeelement än *-s*: *gästabud, prästestånd*).

Ännu mer komplexa är enkla substantiv som slutar på obetonad vokal (främst *-a* eller -e). Det finns minst fyra möjligheter; t.ex. *lampskärm,* där substantivet tappar slutvokalen, *temadag,* där substantivet bevaras i sin helhet, *gatukorsning*, där substantivets slutvokal ersätts med ett fogeelement och *dikeskörning*, där substantivet bevaras, och fogeelementet *-s* sätts in. Dessutom kan samma ord följa fler än ett av de fyra mönstren: *gatukorsning* men *gatlykta.*

Substantiv som själva är sammansättningar och slutar på konsonant har huvudregeln att de får foge-s: jämför *barnboksrea* och *bokrea.* Det finns dock undantag från huvudregeln som är svåra att förklara, t.ex. *tonfisksallad.*

Sammansatta substantiv som slutar på vokal, främst *-a* och *-e,* är kanske svårast att hantera av alla substantiviska förled. Heter det tvättstugetid eller tvättstugstid?

2.1.3 Ihop eller isär?

Jag ska nu återknyta till frågan om när ett ord ska skrivas ihop respektive isär. Enligt Svenska Språknämndens skrivregler ska alla sammansättningar skrivas som ett ord. För det mesta har vi inga svårigheter att avgöra vad som är sammansättning, speciellt inte när det gäller substantiv, adjektiv eller verb. Man kan oftast höra på betoning och intonation om det ska vara ett ord eller flera, jämför *fem ton - femton.*

I vissa fall går det att välja fritt mellan en sammansättning och ett flerordsuttryck, som till exempel *en tvåmeters man, en tvåmetersman.* Undantaget är dock i bestämd form där endast sammansättning kan användas (*tvåmetersmannen*). Observera dessutom att flerordsuttrycket och sammansättningen kan ha olika betydelser; *en grön sak - en grönsak.*

När det gäller fasta uttryck är det ibland tveksamt om de ska betraktas som sammansättningar eller som fasta flerordsuttryck, t.ex. *för resten* eller *förresten.* Detta gäller särskilt vissa fasta prepositionsuttryck där det ofta inte går att höra om det är sammansättningar eller ej.

Det finns vissa tumregler för hur man skriver fasta prepositionsuttryck. Om den starkaste betoningen ligger på det första ordet: skriv ihop. Ex. *alltför, därvidlag.* I övriga fall: skriv isär. Ex. *för all del, i morgon.* Detta enligt Svenska Skrivregler. Den som fortfarande är osäker hänvisas till SAOL som i sin tur ger valfrihet i nästan alla fall där skrivregelsamlingen rekommenderar särskrivning, även för ord som *för all del* och *i morgon*, de som i Svenska Skrivregler lyfts fram som typfall. Enligt SAOL kan man också välja mellan att skriva *i stället* och *istället*, *i fred* och *ifred* liksom *för resten* och *förresten*.

Sammansättningar med flerordsuttryck som förled skrivs för det mesta i ett ord; *svarta börsen + haj = svartabörshaj.* Några fler exempel: *godnattkyss, sanktbernhardshund, rödakorssyster.*

Svenska Skrivregler säger också att bindestreck används då förledet är ett flerordigt egennamn och där alla delar har namnkaraktär och stor bokstav, t. ex. *New York-borna*. När

förleden består av fler än två ord kan bindestreck sättas mellan alla ord. Ex. *berg-och-dalbana, tio-i-topp-lista*.

2.2 Särskrivningar

I många fall är det svårt att veta om ett ord ska skrivas ihop som en sammansättning eller inte. Det hela kompliceras ytterligare av att Svenska Skrivregler och SAOL ibland ger olika vägledning, vilket beror på att Svenska Språknämnden (Svenska Skrivregler) har högre krav på en god och konsekvent stil.

Som vi sett tidigare rekommenderar Svenska Skrivregler ofta sammanskrivning där SAOL anger valfrihet. Detta gäller främst fasta prepositionsuttryck som tycks vara de svåraste fallen. Dessa ord nämns ofta i artiklar där arga insändare irriterar sig på hur framför allt yngre skriver isär t.ex. *i bland*. Själv har jag funnit att majoriteten av särskrivningar man ser är substantiv + substantiv, som till exempel *häst lek sak med äppel smak.*. Alltså de som normalt betraktas som enkla och självklara sammansättningar och där man får ett entydigt svar i SAOL, Svenska Skrivregler etc. I min egen undersökning som beskrivs mer ingående under avsnitt 4.1 nedan är 70% av särskrivningarna av typen substantiv + substantiv. På Internet var det inte särskilt svårt att finna särskrivningar av denna typ. Dessutom fann jag att de som skrivit var ganska konsekventa med att skriva isär sammansättningarna. En felaktig särskrivning ledde oftast till fler.

Många ord kan man tycka ser väldigt underliga ut då de skrivs isär till exempel *lill tån, å mynningar, skid åkning* med flera. Man kan ställa sig frågande till varför sådana ord skrivs isär. Syns det inte att de hör ihop?

En del av särskrivningarna substantiv + substantiv är sådana där ett foge-s sitter på förleden. En trolig förklaring till varför de skrivs isär är att den som skriver tar fel och ser det som ett genitiv-s.

I tryck förekommer särskrivningar ofta i rubriker, på skyltar, på menyer m.m. Med andra ord i text som inte är löpande text. På Webben hittade jag däremot massor av särskrivningar i löpande text. I allmänhet verkar det som om det skrivs mer informellt på nätet vilket kan bero på att vem som helst kan publicera något där. De flesta som skriver på Internet är varken författare, journalister eller copywriters. De har dessutom sällan en korrekturläsare till hjälp.

Ord som är direktöversättningar från andra språk (oftast från engelska) kan också leda till felaktiga särskrivningar. Jag har stött på sådana särskrivningar framför allt i datorsammanhang och i tekniska sammanhang där författaren/översättaren har sett dessa ord på engelska och sedan direktöversatt. Till exempel *home page* blir *hem sida*.

2.3 Språkgranskningsprogrammet Granska

Granskaprojektet (eg. Datorstöd för språklig granskning under skrivprocessen) har pågått under 1996-1997 under två projektperioder. Under den första perioden utvecklades en granskningsfunktion som kunde finna språkliga avvikelser i skrivna texter. Dessa avvikelser var ortografiska (t.ex. *Lena's*), stilistiska (t.ex. *våran*) och grammatiska (t.ex. *ett bil*). Granskningsfunktionen var från början textbaserad och batchorienterad och kördes i terminalfönster under Unix eller MS-Windows. Funktionen hade inget eget lexikon för att tagga texten utan använde två utomstående program för tilldelning av ordklass- och

böjningsinformation. Antalet granskningsregler var begränsat och framför allt saknades ett funktionellt användargränssnitt.

Under den andra projektperioden vidareutvecklades funktionalitet, arkitektur, innehåll och användargränssnitt. Bland annat integrerades granskningsfunktionen i en grafisk skrivmiljö för Windows och byggdes ut med ett internt lexikon och en taggningsfunktion som identifierar ord och förser dem med ordklass- och böjningsinformation.

För taggning använder sig Granska av ett lexikon som är skapat utifrån den ordklassmärkta textdatabasen SUC. Lexikonet har kompletterats med ord och ordformer från andra källor för att Granska lättare ska kunna hitta felaktigheter. De ord som har lagts till har kompletterats med lemma, vilket gör att man får fram ordets grundform utifrån dess böjningsformer.

I Granskasystemet ingår en lista med språkregler som används för att finna eftersökta egenskaper (fel) i texten. Det finns regler för att hitta felaktiga särskrivningar men endast då det gäller fasta fraser som till exempel *ifall, alltför*. Reglerna är direkt tagna från Svenska Skrivregler.

Användaren kan själv välja en regelfil med de regler som ska användas. Reglerna består av logiska uttryck som beskriver egenskaperna hos den följd av ord som programmet ska reagera på. (Användarinstruktion för Granska.[1])

Ex. _VB & _SUP
_VB & _SUP

Regeln ovan hittar två verb i supinum i följd som i den felaktiga meningen ”jag har länge *velat haft* en bil”. (I reglerna används SUC-taggar enligt appendix A.)

Förutom regler i form av taggningsegenskaper kan man också leta efter ordförekomster eller både och, som i exemplet nedan.

Ex. man & _PN

Regeln hittar alla förekomster av ordet *man* när det förekommer som pronomen.

En av bristerna i nuvarande version av Granska är att det inte finns någon dis-ambiguerare. Därför skulle regeln i exemplet ovan vara meningslös. Granska väljer inte mellan alternativet pronomen och alternativet substantiv utan taggar ordet med bägge. Med regeln ovan får man alltså fram alla förekomster av *man* oavsett om de i kontexten är pronomen eller substantiv. En disambiguerare har dock tagits fram och kommer att finnas med i nästa version[2].

Ett annat problem är att många egennamn (ortsnamn, personnamn) inte taggas alls av Granska. Även här löses problemet i nästa version med hjälp av en ordklassgissare med vilken hanteringen av okända ord kan förbättras.

Det språkgranskningssystem som existerar i dag fortsätter nu att utvecklas i språk-teknologiprojektet *Integrerade språkverktyg för skrivande*[3]. Förutom en disambiguerare och en ordklassgissare kommer ett rättstavningsprogram *Stava* (Domeij et al., 1998) att

[1] *Användarinstruktion för Granska*:
http://www.nada.kth.se/kurser/kth/2D1418/laborationer97/granska/Granska.html.
[2] Mer information om taggern finns på http://www.nada.kth.se/theory/projects/granska/tagga.html.
[3] Mer information finns på http://www.nada.kth.se/iplab/langtools/.

integreras i programmet. Granska kommer också att utvecklas med funktioner för lingvistisk sökning och redigering, t.ex. passivisering och topikalisering.

Många av problemen som finns kvar att lösa i utvecklingen av Granska har att göra med det grundläggande problemet med språklig flertydighet. Ett bra exempel på detta är felaktigt särskrivna sammansättningar, till exempel kan två ord både förekomma som en sammansättning och som två separata ord (*sjuksköterska – sjuk sköterska*). Språklig flertydighet av denna sort kan bara lösas fullt ut den dag datorerna har samma kunskap om världen som vi människor har.

3 METOD

3.1 Material

Det var svårt att få fram material om felaktigt särskrivna sammansättningar. För att studera särskrivningar gjorde jag därför en egen undersökning där jag själv samlade ihop en mängd felaktiga särskrivningar. Det fanns även ett behov för utvecklingen av Granska att få fram en felsamling med kontexter som testmaterial för granskningsfunktionen.

Den resulterande felsamlingen består av 389 särskrivningar med kontexter hämtade från Internet. Jag använde mig av sökord som till exempel *dator program* i AltaVista och fick på så vis fram ett antal dokument som innehöll denna särskrivning. Då jag gick igenom varje dokument fann jag nästan i samtliga fall fler meningar i samma dokument som innehöll alla möjliga sorters särskrivningar. Som jag nämnt tidigare så har de flesta författare av felaktiga särskrivningar varit mer eller mindre konsekventa med att skriva isär sammansättningar.

Nackdelen med metoden är att jag kan ha fått ett snedvridet resultat då jag framför allt sökt på särskrivningar av typen substantiv + substantiv. Det stora flertalet särskrivningar jag fick fram var dock inte de jag hade som sökord i AltaVista. Även om just de särskrivningar som jag från början använde som sökord är i minoritet i min felsamling så finns det en risk för att de övriga särskrivningarna är av samma sort eftersom de som skriver kan tänkas ha en tendens att upprepa samma typ av fel.

3.2 Tillvägagångssätt

För att komma på hur man ska kunna detektera felaktigt särskrivna sammansättningar började jag med att analysera och beskriva konstruktionen av olika typer av särskrivningar. Till hjälp hade jag min egen felsamling som taggades med en preliminär version av Granskas nya taggdisambiguerare.

Jag gick igenom mitt taggade material och plockade ut de särskrivna sammansättningarna ur deras kontexter. Jag upptäckte att vissa (75 st.) var feltaggade. För att få ett så stort underlag som möjligt för min typologi rättade jag dessa för hand. De vanligaste felen var gissningsfel, det vill säga fel som berodde på att taggern inte kände till ordet och gissade fel. En del fel berodde också på stavfel från författarens sida som gjorde att taggern inte taggade rätt. För en diskussion om problematiken kring rättningen se avsnitt 5.

Därefter delade jag in särskrivningarna i kategorier efter ordklasstyper i särskrivningens förled och efterled. Jag beräknade frekvens för de olika typerna. På så vis fick jag fram en typologi, helt och hållen baserad på min insamling. Jag kom fram till att den i särklass vanligaste typen av särskrivning var substantiv + substantiv. Det var dessa jag valde att

koncentrera mig på när jag sedan gick vidare med att detektera felaktigt särskrivna sammansättningar. Utifrån mönster jag såg bland särskrivningarna och utifrån egna resonemang skapade jag regler för att med hjälp av Granska finna särskrivningar i en text.

När jag körde dessa regler i Granska utnyttjade jag den inbyggda taggern som i den version jag använde ej hade någon disambiguerare.

4 RESULTAT

Resultatet består av tre delar: en del som beskriver olika typer av särskrivningar och som utgör grunden för den andra delen som i sin tur presenterar förslag på regler i Granska för att detektera felaktiga särskrivningar. Den tredje delen är en sammanfattning av resultat.

4.1 Typologi över olika särskrivningstyper

Med hjälp av min samling av felaktigt särskrivna sammansättningar ville jag få fram en beskrivning av hur olika särskrivningstyper kan se ut. I min typologi har jag delat in särskrivningarna efter ordklasser. Med ordklasser menar jag ordklassen för de ord (för- och efterled) som särskrivningen består av. Jag har även tagit med sammansättningar bestående av tre led som listas separat. Nedan följer en tabell som visar frekvens för varje särskrivningstyp:

Ordklasser	*Antal*	*%*
<nn><nn>	276	70%
<nn><jj>	31	8%
<nn><vb>	3	
<nn><ab>	1	
<nn><pc>	3	
<nn><pm>	1	
<pm><nn>	14	3,6%
<jj><nn>	10	2,6%
<jj><jj>	8	2,1%
<jj><vb>	3	
<vb><nn>	6	1,5%
<pp><nn>	2	
<pp><jj>	1	
<pp><vb>	1	
<pp><pn>	1	
<pp><ha>	1	

Ordklasser	*Antal*	*%*
<pp><pp>	1	
<ab><pl>	1	
<ab><pp>	1	
<ab><jj>	2	
<ab><nn>	1	
<ab><vb>	3	
<ab><pc>	1	
<rg><jj>	1	
<rg><nn>	1	
<pn><nn>	1	
<uo><nn>	1	
Tre led:	*Antal*	*%*
<rg><nn><nn>	3	
<jj><nn><nn>	3	
<nn><nn><nn>	6	1,5%
<nn><jj><jj>	1	
Summa	389	

Tabell 1[4] - Frekvenser för olika typer av särskrivningar

[4] Granska använder de taggar som finns i SUC (Stockholm Umeå Corpus). En beskrivning av dessa finns med i appendix A.

Som framgår av diagrammet är särskrivningar av typen substantiv + substantiv i särklass vanligast. De representerar 70% av de felaktiga särskrivningarna i min felsamling. Man kan också se att substantiv + substantiv även förekommer i de flesta särskrivningar med tre led.

Tabellen nedan visar hur ofta en viss ordklass förekommer i det totala antalet särskrivningar. Observera att summan i tabellen överstiger 100% eftersom siffrorna anger hur ofta en ordklass förekommer i *något* led i särskrivningarna.

Ordklasser	*Antal*	*%*
substantiv	364	94%
adjektiv	60	15%
adverb	10	2,6%
verb	10	2,6%
prepositioner	8	2,1%

Tabell 2 – Frekvenser för förekomst av ordklasser i något led i särskrivningarna

Substantiv ingår i något av leden i sammanlagt 18 av de 31 olika typerna och i 364 av särskrivningarna. Med andra ord ingår substantiv i 94% av särskrivningarna. Det näst vanligast förekommande är adjektiv som finns i 15% av fallen. Det är med andra ord stor skillnad mellan frekvensen för särskrivningar som innehåller substantiv och övriga.

4.2 Detektion av felaktiga särskrivningar

Jag har försökt detektera felaktigt särskrivna sammansättningar genom att skapa regler i Granska. Reglerna bygger på de särskrivna sammansättningarnas ordklasser och övriga morfologiska taggar. Jag har valt att leta efter särskrivningar av typen substantiv/egennamn + substantiv. En beskrivning av hur reglerna formuleras finns under rubriken 2.3 ovan.

Jag provade mig fram med ett antal regler och redovisar nedan de som är mest illustrativa och/eller relevanta. Reglerna tillämpades på min samling meningar innehållandes felaktiga särskrivningar. I det här skedet användes Granskas egen taggningsfunktion. I den versionen jag använde saknades både disambiguerare och en gissningsfunktion som gissar taggar på ord den inte känner igen. Texten som användes som input var inte heller stavningskontrollerad (rättad). Nedan följer ett exempel på en mening taggad med den inbyggda taggern:

Osäkra JJ.POS.UTR/NEU.SIN.DEF.NOM | JJ.POS.UTR/NEU.PLU.IND/DEF.NOM
på AB | PL | PP
ifall SN
de PN.NEU.SIN.DEF.SUB/OBJ | PM.NOM | UO | PN.UTR/NEU.PLU.DEF.SUB | DT.UTR/NEU.PLU.DEF
kan VB.PRS.AKT
slå VB.IMP.AKT | VB.INF.AKT
i PN.UTR/NEU.PLU.IND.SUB/OBJ | AB | PM.NOM | NN.NEU.SIN.IND.NOM | UO | RG.NOM | PL | PP
sönder AB | PP
fäljen

på AB | PL | PP
cykel NN.UTR.SIN.IND.NOM
hjulen NN.NEU.PLU.DEF.NOM
rullar VB.PRS.AKT
vi PM.NOM | RG.NOM | PN.UTR.PLU.DEF.SUB
sakta JJ.POS.UTR/NEU.SIN/PLU.IND/DEF.NOM | VB.INF.AKT | AB.POS
fram AB | PP
genom PL | PP
den PN.UTR.SIN.DEF.SUB/OBJ | DT.UTR.SIN.DEF
grådisiga
morgonen NN.UTR.SIN.DEF.NOM

I exempelmeningen ser man att det finns fler än ett taggningsalternativ för vissa ord. Till exempel kan man se att ordet ”i” har taggats som pronomen, adverb, egennamn, substantiv, utländskt ord, räkneord (romerskt), partikel och preposition! Man kan också se att vissa ord inte har taggats alls som till exempel ”fäljen” (som är felstavat) och ”grådisiga” som taggern inte känner igen.

4.2.1 Regler i Granska

Jag redovisar reglerna nedan genom att för varje regel ange ett exempel, bakgrund, resultat och slutsats. Under bakgrund redogör jag kort för hur jag kom fram till just den regeln.

Regel 1
(Letar efter ett egennamn i genitiv följt av ett substantiv i bestämd form).

_PM & _GEN
_NN & _DEF

Exempel:
Tunabergs veteranen
Bakgrund:
Jag upptäckte några särskrivningar i min taggade felsamling där förledet var ett egennamn med foge-s på slutet och där efterledet var ett substantiv i bestämd form.

Strategin med regeln är att Granska ska tagga förledet som om det vore ett egennamn i genitiv. I allmänhet följs ett egennamn i genitiv oftast av ett substantiv i *obestämd* form, (till exempel *Johans bil).* Att egennamn i genitiv följs av ett substantiv i *bestämd* form förekommer förmodligen bara i samband med felaktigt särskrivna sammansättningar.
Resultat:
Tyvärr gav den här regeln inga träffar i min felsamling eftersom många egennamn inte taggas alls av Granska.

Funna felaktiga särskrivningar: 0
Totalt antal träffar: 0

Slutsats:
Den här regeln angriper inte en särskilt stor andel av alla särskrivna sammansättningar. Däremot borde regeln ha hög precision, det vill säga de ord den hittar är med hög sannolikhet särskrivna sammansättningar. Med en bättre taggning hade denna regel gett bra resultat.

Regel 2
(Letar efter ett substantiv i genitiv följt av ett substantiv i bestämd form).

_NN & _GEN
_NN & _DEF

Exempel:
ursprungs instruktionen
Bakgrund:
Jag upptäckte några särskrivningar i min taggade felsamling där förledet var ett substantiv med foge-s på slutet och där efterledet var ett substantiv i bestämd form.

Strategin med regeln är precis som i föregående regel att Granska ska tagga förledet som om det vore ett substantiv i genitiv. I allmänhet följs ett substantiv i genitiv oftast av ett substantiv i *obestämd* form, (till exempel *blommans färg).* Att ett substantiv i genitiv följs av ett substantiv i *bestämd* form förekommer förmodligen bara i samband med felaktigt särskrivna sammansättningar.
Resultat:
Regeln gav två träffar som bägge var felaktiga särskrivningar. Felsamlingen innehöll många fler som inte upptäcktes på grund av brister i taggningsfunktionen.

Funna felaktiga särskrivningar: 2
Totalt antal träffar: 2

Slutsats:
Den här regeln angriper en något större andel av alla särskrivna sammansättningar jämfört med regel 1. Regeln har hög precision, det vill säga de ord den hittar är med hög sannolikhet särskrivna sammansättningar. Med en bättre taggningsfunktion hade denna regel gett fler träffar.

Regel 3
(Letar efter ett substantiv i nominativ följt av ett substantiv i bestämd form).

_NN & _NOM
_NN & _DEF

Exempel:
cykel hjulen

Bakgrund:
Många av särskrivningarna i min taggade felsamling bestod av ett substantiv i nominativ som förled och ett substantiv i bestämd form som efterled.

Strategin med regeln är att det sällan står ett substantiv i nominativ före ett substantiv i bestämd form förutom i felaktigt särskrivna sammansättningar.

Resultat:
Regeln gav 33 träffar som var särskrivningar och 41 felaktiga träffar som inte var särskrivna sammansättningar. De senare har jag delat in i olika kategorier beroende på orsaken till felet. 36 av de 41 felaktiga träffarna orsakades av bristen på disambiguering (nedan kallas de ”disambigueringsfel”) där det första ordet oftast var prepositionen ”*i*” som taggats som substantiv. 4 av träffarna var par av substantiv som gav träff trots att det fanns ett kommatecken mellan dem. En av de felaktiga träffarna kom sig av ett annat fel som författaren till texten gjort.

Funna felaktiga särskrivningar: 33
Disambigueringsfel: 36 (första ordet oftast ”i”)
Annat fel (från författarens sida): 1
Kommateckenfel: 4
Totalt antal träffar: 74

Slutsats:
Den här regeln lyckas bra om man får bort disambigueringsfelen. Ingen av de felaktiga träffarna beror på regeln i sig. Regeln har hög precision, det vill säga de ord den hittar är med hög sannolikhet särskrivna sammansättningar.

Regel 4
(Letar efter ett substantiv förutom ”i” följt av ett substantiv i bestämd form).

_NN & ^ i
_NN & _DEF

Exempel:
cykel hjulen
Bakgrund:
Regeln är en sammanslagning av regel 2 och 3 med tillägget att jag plockat bort alla ordpar där första ordet är ”i” eftersom Granska taggar ”i” bland annat som ett substantiv. Det första substantivet kan alltså stå i antingen nominativ eller genitiv.

Resultat:
Regeln gav 35 träffar som var felaktiga särskrivningar (33+2). Den gav 17 träffar som inte var felaktiga särskrivningar enligt nedan (se även regel 3):

Funna felaktiga särskrivningar: 35
Disambigueringsfel: 12
Annat fel: 1
Kommateckenfel: 4
Totalt antal träffar: 52

Slutsats:
Genom att få bort ”i” blev jag av med en stor del av felen som beror på bristen på disambiguerare. Ingen av de felaktiga träffarna beror på regeln i sig. Regeln har hög precision, det vill säga de ord den hittar är med hög sannolikhet särskrivna sammansättningar. Med disambiguerad taggning torde den här regeln bli effektiv.

Regel 5

(Letar efter en determinerare i ett visst genus följt av ett substantiv i ett annat genus som i sin tur följs av ett substantiv i determinerarens genus).

_DT & _UTR
_NN & _NEU
_NN & _UTR

Exempel:
en hus nyckel

_DT & _NEU
_NN & _UTR
_NN & _NEU

Exempel:
ett cykel ställ
Bakgrund:
Strategin med regeln är att använda kongruensen mellan determinerare och följande ord. Om en determinerare följs av ett substantiv med ett annat genus som i sin tur följs av ett substantiv med samma genus som determineraren är det med stor sannolikhet en felaktigt särskriven sammansättning.
Resultat:
Tyvärr gav den här regeln bara två träffar i min felsamling. Den ena var en felaktigt särskriven sammansättning och den andra berodde på bristen på disambiguerare.

Funna felaktiga särskrivningar: 1
Disambigueringsfel: 1
Totalt antal träffar: 2

Slutsats:
Den här regeln angriper inte en särskilt stor andel av alla särskrivna sammansättningar. Däremot borde regeln ha hög precision, det vill säga de ord den hittar är med hög sannolikhet särskrivna sammansättningar.

4.3 Sammanfattning av resultat

Jag har presenterat en typologi över olika särskrivningstyper baserad på förledens och efterledens ordklasser. Jag har sedan försökt detektera särskrivningar av typen substantiv + substantiv som var den i särklass mest förekommande i min typologi. Reglerna som jag redovisar ovan har alla hög precision men täcker bara in en del av särskrivningarna av typen substantiv + substantiv. Regel 1 koncentrerar sig på egennamn i förledet. Den regel som lyckades bäst var regel 4 som är en utveckling och sammanslagning av regel 2 och 3. Regel 4 täcker in sammansättningar där ett substantiv följs av ett substantiv i bestämd form. Denna typ av särskrivningar utgjorde 23 procent av det totala antalet ”substantiv + substantiv” i min felsamling. I resten av fallen hade särskrivningarna efterled i obestämd form. Det hade varit önskvärt att hitta en metod för att detektera dessa. Det underliggande problemet är att två substantiv i följd där det andra substantivet står i obestämd form förekommer relativt ofta i text utan att det för den skull är en felaktig särskrivning. Regel 5 angriper i och för sig även särskrivningar där efterledet är i obestämd form men den är väldigt smal (den ger få träffar).
I en implementation skulle man kunna ha med reglerna 1, 4 och 5.

Sammanfattningsvis kan man säga att med den metod jag har valt kan man endast angripa en begränsad mängd felaktiga särskrivningar, dock med hög precision hittills.

5 DISKUSSION

Ett problem jag stötte på när jag skulle räkna frekvenser för olika särskrivningstyper var taggningen av särskrivna sammansättningar. Grundproblemet med taggning av felaktiga särskrivningar är att man taggar något som från början är fel. Egentligen är ju en särskriven sammansättning *ett* ord fast felskrivet. Det kan vara så att förledet inte är ett ord i sig för att det har tappat en vokal på slutet (till exempel *skid* i *skid åkning)* och/eller för att det har tillkommit ett fogeelement (till exempel *åsne* i *åsne taxi).* Om fogeelementet är ett *-s* tolkar taggern ofta förledet som ett substantiv i genitiv vilket utnyttjas av mina regler ovan. En annan orsak till felaktig taggning är stavfel gjorda av författaren.

Taggern som användes på min felsamling, när jag skulle ta fram min typologi, hade en ordklassgissare. Den försökte tagga även okända förled och felstavade ord vilket gjorde att vissa ord blev feltaggade. Jag hade två alternativ: att stryka feltaggade ord ur min undersökning eller att tagga om dem manuellt. Om jag strök orden skulle det leda till ett mindre underlag för min typologi. Därför valde jag det andra alternativet och rättade alla feltaggade särskrivningar för hand. Ibland hade jag problem med att rätta taggningen; vad är *villo* i *villo vägar* till exempel? Är *superskalär* i *superskalär processor* ett substantiv eller ett adjektiv? I de fall där förledet inte är ett ord har jag valt att tagga om det som det ord

som ligger närmast till hands, till exempel har jag taggat om *grott* i *grott typ* från verbet *gro* till substantivet *grotta*.

Samma problem med förled och stavfel uppkom även när jag körde mina regler i Granska. Då använde jag den inbyggda taggern som varken hade disambiguerare eller ordklassgissare. Detta innebar att förled som inte är ord samt felstavade ord inte taggades alls vilket i slutändan ledde till att reglerna missade särskrivningar.

Jag valde att köra reglerna på min egen felsamling vilket är en begränsning eftersom jag delvis utgick från min felsamling, både medvetet och omedvetet, när jag skrev reglerna. Anledningen till detta val var att jag inte hade någon representativ text som innehöll tillräckligt många felaktiga särskrivningar för att det skulle bidra till resultatet.

5.1 Framtida arbete

Avslutningsvis kan man säga angående resultatet att det vore intressant att köra mina regler i den nya versionen av Granska där en disambiguerare och en ordklassgissare finns med. I ett fortsatt arbete kan man också studera andra typer av särskrivningar än substantiv + substantiv.

Vidare vore det intressant att studera korrektion av felaktigt särskrivna sammansättningar, dels hur det ska implementeras och dels hur man löser problematiken kring användargränssnittet. Det gäller att se till att användaren får hjälp av datorn.

Det går också att angripa problemet med särskrivna sammansättningar till exempel genom att redan i tokeniseringsstadiet detektera typiska förled (*åsne* i *åsne taxi*). Exempelvis skulle Granska kunna tagga sådana direkt som element i en sammansättning.

6 REFERENSER

Användarinstruktion för Granska:
http://www.nada.kth.se/kurser/kth/2D1418/laborationer97/granska/Granska.html

Domeij, R., Knutsson, O., Larsson, S., Severinson Eklund, K. & Rex, Å., 1998. *Granskaprojektet 1996-1997*. IPLab, Nada, KTH, Stockholm.

Domeij, R., Knutsson, O. & Larsson, S., 1996. *Datorstöd för språklig granskning under skrivprocessen: en lägesrapport*. IPLab, Nada, KTH, Stockholm.

Malmgren, Sven-Göran, 1994. *Svensk lexikologi*. Studentlitteratur, Lund.

SvD 971207. Strömquist, Siv, *Leverpastejen eller lever pastejen?* Artikel i Svenska Dagbladet, Stockholm.

Svenska Akademiens ordlista över svenska språket, (SAOL), 1986. Norstedts förlag, Stockholm.

Svenska skrivregler utgivna av Svenska språknämnden, 1991. Almqvist & Wiksell, Stockholm.

Thorell, Olof, 1981. *Svensk ordbildningslära*. Esselte Studium, Stockholm.

APPENDIX A: SUC-TAGGAR OCH STILTAGGAR

SUC-tagg	Betydelse	Exempel på ord som har taggen
AB	adverb	inte
AKT	aktiv form	spelar
AN	förkortning	t.ex.
DEF	definit	bilen
DL	skiljetecken	.
DT	determinerare (artikel)	den
GEN	genitiv	bilens
HA	frågande/relativt adverb	när
HD	frågande/relativ determinerare	vilken
HP	frågande/relativt pronomen	vem
HS	frågande/relativ possessiv	vems
IE	infinitivmärke	att
IMP	imperativ	spring
IN	interjektion	ja
IND	indefinit	bil
IND/DEF	indefinit/Definit	gula
INF	infinitiv	spela
JJ	adjektiv	gula
KN	konjunktion	och
KOM	komparativ	gulare
KON	konjunktiv form	vare
MAD	skiljetecken i slutet av en mening	.
MAS	maskulinum	gule
MID	skiljetecken inom en mening	,
NEU	neutrum	huset
NN	nomen (Substantiv)	bilen
NOM	nominativ	bilen
OBJ	objektform	mig
PC	particip	kastad
PL	partikel	om
PLU	plural	bilar
PM	egennamn	Svensson
PN	pronomen	hon
POS	positiv	gul
PP	preposition	till

PRF	perfekt	kastad
PRS	presens	spelar
PRT	preteritum	spelade
PS	possessiv	mina
RG	räkneord grundtal	två
RO	räkneord ordningstal	andra
SFO	s-form, passiv eller deponens	behövdes
SIN	singular	bil
SIN/PLU	singular/plural	boxande
SMS	sammansättning	pojk- (och flickrum)
SN	subjunktion	om
SUB	subjekt	jag
SUB/OBJ	subjekt/objekt	den
SUP	supinum	spelat
SUV	superlativ	gulast
UO	utländskt ord	action
UTR	utrum	bil
UTR/NEU	utrum/neutrum	gula
VB	verb	spela

Stiltagg	**Betydelse**	**Exempel på ord som har taggen**
ABSO	abstrakt ord	rubricerade
ABSP	abstrakt preposition	beträffande
DATA	dataterm	password, e-mail
DATF	dataförkortning	CD-ROM, kBit
FELT	Ord som lätt feltolkas	bordla
FOAL	formella eller ålderdomliga ord	förebringa, besvärstalan
FRAM	främmande ord	stigmor
FSMS	felaktig sammansättning	nuförtiden, varsin
ONFL	onödiga förlängningar	befrämja
PAVB	partikelverb	handha
SVBA	svårbegripligt ord, mer än 75 % missförstår.	delegerar
SVBB	svårbegripligt ord, mellan 50 och 75 % av svenskarna missförstår	disponibel
TIDU	tidsuttryck	julafton
TVET	tvetydiga ord	utgick
VARD	vardagliga ord	dej

APPENDIX B: UTDRAG UR FELSAMLING

141. En medel bulgar tjänar ca. 600 - 800 sek / månad.

142. Vi cyklar snabbt vidare i vetskap om att vi bör vara inom hus då mörkret inträffar.

143. Inte en fågel sång hörs.

144. Landskapet är som en enda lång berg och dal bana.

145. Russi som mannen heter är jätte vänlig.

146. Han hör till de som har det bättre ställt och äger en vetemjöls fabrik.

147. Under vår utsökta middag bestående utav hönsoppa , pitka bröd (som är ett dubbel vikt bulgariskt typ pizza bröb fyllt med härlig färsk får och get ost ,så gott att det smälter i munnen) och Russis eget hemgjorda plommon vin , som får det att köra till ordentligt i magarna på oss , berättar Russi om sitt land.

148. Till och med hästar är för dyra ,svarar han på min fråga varför vi bara sett åsnor som drag djur.

149. Han och hans fru var på besök hos en god vän i hamnstaden Varna.Bilen var låst, larmad och klockan var lunch tid.

150. Mycket ledsna och bedrövade begav de sig till polis stationen.

151. Det var en nästan helt ny wolkvagns buss och försäkringen betalade inte ens ut hälften av bilens värde.

152. Cykel hälsningar Acke och Boel.

153. Ralli arbetar som så många andra med skandinaviska turister på sommarhaövåret, längs Svarta havs kusten.

Ref. 132-153 http://www.outdoor.se/artiklar/cykelresan/rapport6.htm

154. Ayran är en youghurt drink bestående av youghurt uppblandat med källvatten, ibland spetsat med lite salt.

155. Medelhavets turkosblå gröna vågor slår upp mot klipporna och blir till ett gräddvitt skum.

156. Vatten bufflar rullar sig välmående i lervällingen på åkrarna.

157. Vallhundarna bär de traditionella halsbanden med taggiga järn nitar utåt.