Ur Språk & stil NF 16, 2006

Vad kan statistik avslöja om svenska sammansättningar?

Av JONAS SJÖBERGH och VIGGO KANN

I många språktekniska datortillämpningar är det värdefullt att kunna dela upp sammansatta ord i deras beståndsdelar. Exempelvis får man fler träffar vid informationssökning om även dokument som bara innehåller sökordet som förled i en sammansättning hittas (Dalianis 2005). Andra tillämpningar är automatisk avstavning, där man helst vill ha avstavningar gjorda i sammansättningsgränsen, och maskinöversättning, där översättningslexikonet förmodligen inte innehåller så många sammansättningar, men däremot sammansättningens beståndsdelar.

I denna artikel visar vi hur väl ett datorprogram med statistiska metoder (till skillnad från metoder baserade på regler konstruerade av en lingvist) kan komma fram till hur de sammansatta ord som förekommer i en text bör delas upp.

Många sammansättningar har mer än en möjlig tolkning. Ibland finns flera rimliga tolkningar, som »glas-skål» och »glass-skål» (men förmodligen inte »glass-kål») eller »bil-drulle» och »bild-rulle». I de flesta fall är det lätt för en mänsklig läsare att från kontexten förstå vilken tolkning som avses. För automatiska metoder är detta dock mycket svårare, och dessa kan även få problem med tolkningar som för en människa är uppenbart osannolika, till exempel »Ko-rea-kriget».

Det enklaste och vanligaste sättet att dela upp sammansatta ord är att ha ett lexikon där man talar om hur varje ord bör delas upp, vilket förstås inte löser problemet för genuint tvetydiga sammansättningar, men dessa är relativt ovan-

Detta arbete har finansierats av Vinnova, Vetenskapsrådet och KTH. Tack till Svenska Akademien för att vi fått tillgång till SAOL 11 och till Stockholms universitet för att vi fått tillgång till SUC. Tack också till Språk och stils anonyma referenter som har haft värdefulla synpunkter på artikeln.

Artikelförfattarnas kontaktuppgifter: Skolan för datavetenskap och kommunikation, KTH; jsh@nada.kth.se, viggo@nada.kth.se

liga. Ett annat angreppssätt som använts är undersökning av bokstavsföljder som bara kan förekomma i sammansättningar, till exempel »kk», för åtminstone där går en ordledsgräns. För svenska har både regelbaserade metoder (till exempel Karlsson 1992, Dura 1998) och statistiska metoder (till exempel Kokkinakis & Johansson Kokkinakis 1999, Sjöbergh & Kann 2004) använts för automatisk uppdelning av sammansatta ord. Liknande arbeten har gjorts för andra språk, bland annat tyska (Koehn & Knight 2003), norska (Johannessen & Hauglin 1996) och koreanska (Yoon 2000).

Här kommer vi att se på några statistiska metoder för tolkning av sammansatta ord. Tonvikten ligger på valet mellan olika föreslagna tolkningar, men först några ord om hur dessa förslag tillkommer.

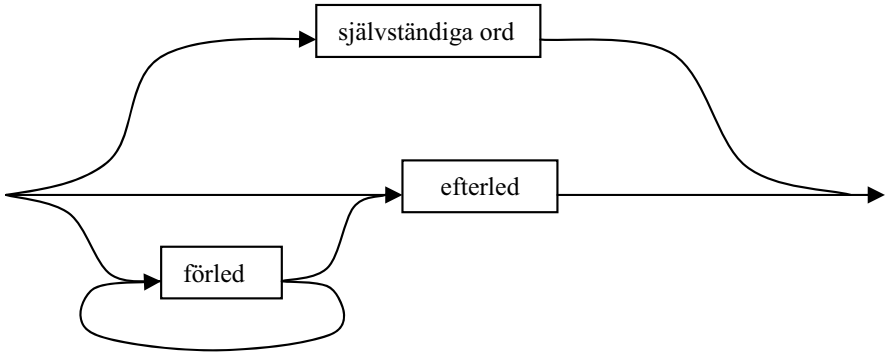
1. Att hitta möjliga tolkningar

För att ta fram möjliga tolkningar av sammansatta ord har vi använt en modifierad version av rättstavningsprogrammet Stava (Domeij m.fl. 1994). Eftersom ett rättstavningsprogram för svenska måste kunna hantera sammansatta ord för att bli användbart finns redan en komponent för detta. Stava använder tre olika ordlistor:

1. *självständiga ord*, innehåller 1 400 ord som inte kan förekomma som delar av sammansättningar, till exempel »inte»
2. *efterledsordlistan*, innehåller 103 000 ord (även vanliga böjningsformer) som kan avsluta sammansatta ord eller fungera som självständiga ord, till exempel »lampor», »stolpen» och »gata»
3. *förledsordlistan*, innehåller de ordformer som kan förekomma först i, eller i mitten av, sammansatta ord, till exempel »gatu», »telefon» och »ladu», totalt 23 000 ordformer.

När ett ords stavning ska kontrolleras används de olika ordlistorna enligt schemat i figur 1. I de enkla fallen hittas ordet direkt i *efterledsordlistan* eller listan över *självständiga ord*. Om ordet är sammansatt slås alla möjliga efterled upp tills ett som finns i *efterledsordlistan* hittas, med början på det längsta möjliga efterledet. Därefter slår programmet upp förledet i *förledsordlistan* för att kontrollera att det är ett riktigt förled. Om förledet också är sammansatt slås delarna upp rekursivt i *förledsordlistan*.

Foge-s, som i »fot-boll-*s*-lag», tillåts i tre- och flerledssammansättningar.



Figur 1. Uppslagningsschema för analys av sammansatta ord.

Däremot finns det inget krav på foge-s i treledssammansättningar, eftersom foge-s inte ska sättas ut då andra och tredje ledet är hårdare knutna till varandra än första och andra ledet (till exempel »skolbokhylla» som är en bokhylla för skolor och inte en hylla för skolböcker; se sida 49–52 i Språkriktighetsboken 2005). I de fall enkla förled bildar sammansättningar med foge-s ingår de i förledsordlistan (till exempel »kvälls»).

Finns både förled och efterled i ordlistorna är ordet godkänt och sökandet avbryts, annars följer nästa möjliga efterled på samma sätt osv. Om ingen möjlig uppdelning hittas rapporteras ordet som felstavat.

I vår tillämpning har vi modifierat algoritmen så att den får producera samtliga möjliga tolkningar, i stället för att avbryta så fort en acceptabel tolkning hittats. Även att hitta samtliga tolkningar går mycket snabbt med metoden ovan. På en vanlig dator kan man analysera över 60 000 ord per sekund.

2. Utvärdering

Vi har utvärderat metoder för uppdelning av sammansättningar genom att för hand gå igenom alla ord programmet delar upp i en text på 50 000 löpord. Texten är tagen från Stockholm-Umeå Corpus, SUC (Ejerhed m.fl. 1992) och består av skriven svenska från olika genrer. Uppdelningsprogrammet delade upp 3 400 sammansatta ord i texten (3 098 unika ordformer), varav 1 200 hade mer än ett förslag. Om inget annat anges syftar resultatsiffror längre fram i artikeln

på hur bra metoderna klarade sig på de 1 200 ord som på detta sätt är tvetydiga. Denna utvärderingstext har bara använts för utvärdering och har alltså inte till exempel använts för att inhämta statistik till de olika metoderna nedan.

Hur många ordled dessa 3 400 sammansättningar består av framgår av tabell 1.

Tabell 1. Fördelning av antal ordled hos de 3 400 sammansättningarna i utvärderingstexten.

Antal ordled	Andel
2	92 %
3	5,0 %
4	3,0 %
5	0,06 %

En del sammansatta ord i texten delades inte alls upp av programmet. Ett manuellt kontrollerat stickprov ord visade att mindre än 1 % av sammansättningarna missades på detta sätt. Bland annat har programmet svårt för sammansättningar som innehåller namn, som »Venusmålningarna» och »Hitchcocklärroboken», eftersom dessa namn inte finns med i förledsordlistan.

Detta problem kan dock lätt minskas genom att man lägger till vanliga namn i *förledsordlistan*. En generellare men något riskfylld lösning är att ord som betraktas som felstavade av Stava får genomgå en extra undersökning, där ett så långt efterled som möjligt försöker hittas i ordet, utan att något matchande förled finns i förledsordlistan. Denna metod skulle göra den korrekta uppdelningen »Venus-målningarna», men den skulle också föreslå uppdelningen »Gotland-skaka» snarare än »Gotlands-kaka». I denna artikel har vi inte använt denna generellare hantering av sammansättingar med namn.

För 99 % av de 2 200 ord som bara hade ett förslag var detta förslag korrekt, och för 99 % av de 1 200 ord som hade flera förslag var ett av dessa korrekt. Här berodde större delen av de problematiska fallen (1 %) också på namn, som i »Ko-rea-kriget» eller »Vig-gen-planen».

Även många ord som inte är sammansatta kan delas upp av programmet, som »vi-ta», »in-gen» och »Ja-mes», men detta har undvikits genom att uppdelning aldrig görs av ord som redan finns i *efterledsordlistan* eller listan med *självständiga ord*. Det förekommer förstås fortfarande att en del ej sammansatta ord analyseras som om de vore sammansättningar. Detta beror på att ordlistorna av naturliga skäl inte innehåller samtliga möjliga ord i det svenska språket, och de ord som inte finns med i ordlistorna men som kan analyseras som sammansättningar antas då vara sammansatta ord. I utvärde-

ringstexterna är det arton ord som delas upp trots att de inte är sammansatta, se tabell 2. Dessa ord räknades inte med i utvärderingarna av metoder för val av tolkning.

Tabell 2. Ej sammansatta ord som ändå delas upp av programmet under generering av analysförslag.

Ord	Analys	Ord	Analys
BIOSets	bi|osets	Demodokos	demo|dok|os
Long-projektets	lo|ng-projektets	Roslin-porträtt	ros|lin-porträtt
Sorghum	sorg|hum	Temo-mätning	te|mo-mätning
Tommys	tom|mys	apropos	ap|rop|os
fulleren	full|leren	instrumentalis	instrumental|is
knappet	knapp|pet	kommunalarna	kommun|alarna
oklassiska	ok|lass|siska	släktforsnig	släkt|fors|nig
silliga	sill|liga	styckena	stycklena
stoans	sto|ans	passagerar-	pass|agerar-

Även en del ord som kan betraktas som sammansatta ignorerades, främst namn som »Hall-ström» och »Alm-gren». Det bedömdes vara mindre intressant att låta dessa ord ingå i studien. Samtliga ord som ignorerades hade endast ett tolkningsförslag, vilket vanligtvis var den korrekta analysen om ordet vore att betrakta som sammansatt.

Sammanfattningsvis har programmet för att hitta möjliga tolkningar en täckning på 99 %. Endast ett fåtal sammansättningar saknar den korrekta analysen bland de föreslagna tolkningarna eller blir helt utan analys. När det gäller precision så är denna också 99 % i bemärkelsen att 99 % av de ord som analyseras har den korrekta analysen som ett av förslagen, detta även inräknat de ord som inte är sammansatta och alltså inte kan ha någon korrekt analys. Detta innebär förstås att de metoder som ska välja korrekt tolkning inte kan få mer än högst 99 % rätt, eftersom de alltid väljer endast bland de föreslagna tolkningarna.

Det finns även andra system för att analysera svenska sammansättningar. Det mest kända systemet är förmodligen SWETWOL (Karlsson 1992), som till skillnad från Stavas enkla lexikonuppslagningsmetod bygger på manuellt skrivna regler för bland annat morfologiska förändringar vid sammansättningar. Precis som uppdelningsprogrammet byggt på Stava levererar SWETWOL samtliga möjliga anlyser av ett ord.

SWETWOL hanterar en del ord som inte Stava klarar, till exempel »Venusmålningarna» och »Viggenplan», medan ord som »Hitchcocklåroboken» inte hanteras av något av systemen. Det finns förstås också ord som Stava hanterar

som inte SWETWOL analyserar korrekt. Av de 3 400 ord uppdelningsprogrammet delade upp i testmaterialet har 94 % den korrekta tolkningen bland de förslag SWETWOL genererar. När korrekt tolkning inte föreslås beror det i princip alltid på att ordet inte analyserats som en sammansättning. Exempel på sådana ord är »yngelboet», »gudaborgen», »nikotingula», »läsvane» och »åskblå». SWETWOL har alltså något högre precision, i princip 100 %, men betydligt sämre täckning.

3. Statistiska metoder för val av uppdelning

Här följer en presentation av ett antal olika metoder för val av rätt tolkning bland flera förslag på uppdelning av en sammansättning. En sammanställning av hur bra dessa fungerar ges till sist i tabell 3.

3.1 En undre gräns, antalet ordled

En enkel metod att välja tolkning är att ta det förslag som har lägst antal ordled, alltså »mun-vinklarna» snarare än »mun-vin-klarna». Denna metod är statistisk i den bemärkelsen att sammansatta ord med två ordled är markant vanligare än sammansättningar med tre ordled, vilka i sin tur är vanligare än de med fyra osv. Man kan även ge en lingvistisk motivering, och metoden har använts i regelbaserade system också (Karlsson 1992).

Vi har använt denna metod som en undre gräns. Om det finns flera förslag med lika få ordled väljs det med längst efterled, motiverat av att det är vanligare med böjningsändelser och avledningar i slutet än med korta efterled, »upp-rättar» snarare än »upprätt-ar». Denna enkla och snabba metod fungerar tämligen bra och väljer rätt förslag i 93 % av fallen i vår testkorpus.

3.2 Ord i kontexten

Sammansättningar med flera rimliga tolkningar kräver förstås att man tittar på kontexten för att förstå vad som avses med ordet just här. Nu är den typen av tvetydigheter mycket ovanliga, men att titta på kontexten kan ge stor hjälp även i andra fall. Ett enkelt sätt att ta hjälp av kontexten är att se om ordleden från något förslag förekommer som självständiga ord eller i någon annan samman-

sättning i närheten. Står det »en bildrulle körde sin bil genom staketet» kan vi se ordet »bil» i kontexten som stöd för tolkningen »bil-drulle».

För att utvärdera hur bra detta fungerar prövade vi följande metod: för varje tolkning räknades förekomster av dess ordled som antingen självständiga ord eller ordled i andra sammansättningar i kontexten. Ett fönster på femtio ord före sammansättningen och femtio ord efter användes. Förekomsterna viktades med avståndet till sammansättningen, så att ord som stod närmare sammansättningen fick högre vikt. Den tolkning som fick högst medelvärde på sina ordled valdes sedan.

Denna metod fungerar dåligt, och gav rätt tolkning i endast 75 % av fallen. Orsaken är glesheten i texten; det är sällan något av ordleden förekommer i närheten. Om man använder ordledens stammar och stammarna för orden i kontexten i stället för själva ordleden går det något bättre, även om det i vissa fall inte är helt klart vad stammen till ett förled är.

Ett stort problem med denna metod är att det är vanligt att felaktiga tolkningar av sammansättningar innehåller många korta ordled (»han-del-s-minister»). Just de korta ordleden har stor sannolikhet att förekomma i kontexten av en slump, eftersom korta ord oftast också är vanliga ord. Om man först sållar bort alla förslag till tolkningar som har fler ordled än det med minst antal ordled, alltså förslag med många korta ordled, och sedan använder ord i kontexten som ledtråd för att välja mellan de förslag som finns kvar fungerar metoden mycket bättre, 94 % rätt.

En människa använder naturligtvis mer information från kontexten än den att ordled förekommer som egna ord i närheten. Till exempel föredrar man tolkningen »bild-rulle» i texter om fotografering. Denna typ av kunskap är svår att fånga automatiskt, men det finns statistiska metoder som försöker göra detta. En sådan är Random Indexing, RI (Sahlgren 2005). I korthet kan man säga att det är ett sätt att mäta hur relaterade två ord är, genom att föra statistik över hur ofta olika ord samförekommer. Ord som samförekommer med ungefär samma ord betraktas som relaterade.

Om man i stället för att som ovan räkna förekomster i kontexten för varje ordled räknar ut hur relaterade ordleden är till orden i kontexten går det ytterligare lite bättre, 96 % rätt, vilket gör metoden till en av de bästa vi undersökt. Det går dock inte bra om man inte först avfärdar alla tolkningar med många korta ordled. Behåller man dem går det i stället riktigt dåligt, 51 % rätt, bland annat för att de korta ordleden ofta är vanliga ord som på så sätt räknas som relaterade till väldigt många ord.

Metoden har också den intressanta egenskapen att den väljer olika tolk-

ningar för samma ord i olika kontexter. I vår undersökning är detta den enda metod som föreslog olika tolkningar för samma sammansättning på olika platser i utvärderingstexten, bland annat ordet »kol-atom», som på ett ställe blev »kola-tom». Tyvärr gällde texten en »kol-atom» även på denna plats, men det är i alla fall en intressant egenskap att kunna föreslå olika tolkningar beroende på kontext. De flesta metoder som använts utnyttjar inte kontexten och föreslår alltså alltid samma tolkning.

3.3 Relation mellan ordleden

Om man väl har ett sätt att mäta hur relaterade ord är kan man använda denna till att se hur relaterade till varandra ordleden i ett förslag är. När vi tittar på »glasskål» kan vi fråga oss om man brukar ha glass i skålar, ha skålar av glas eller äta glass tillsammans med kål.

Vi prövade detta genom att för varje förslag beräkna medelvärdet av hur relaterade ordleden var till varandra, enligt RI-metoden. Liksom i förra avsnittet går det riktigt dåligt att skilja ut förslag med många korta ordled, 46 %, och riktigt bra om man först sållar bort förslag med många ordled, 96 %.

3.4 Ordledsfrekvenser

En metod att utnyttja statistik för att tolka sammansatta ord är följande: om vi inte har något annat att gå på så föredrar vi »glas-skål» framför »glass-skål», eftersom det är vanligare med sammansättningar med »glas» än med »glass». För att pröva detta samlades statistik över frekvenser för förled och efterled. Det förslag som hade högst geometriskt medelvärde för sina ordleds frekvenser valdes sedan som tolkning.

Denna metod fungerade ungefär lika bra som att bara titta på antalet ordled, 93 % rätt. Återigen är glesheten i språket ett problem. Många av de föreslagna ordleden förekom över huvud taget inte i de textmängder statistiken hämtats från. Flera olika textmängder användes som korpus för denna statistik: de en miljon orden i SUC (alltså inte frekvenser över ordleden, eftersom sammansättningar där inte är uppdelade); 84 000 sammansättningar ur SAOL 11; de 84 000 sammansättningarna, men med frekvenser för dem beräknade från SUC; 300 sammansättningar från andra texter i SUC än dem vi utvärderat, handannoterade med sina korrekta uppdelningar.

De bästa resultaten uppnåddes med sammansättningarna från SAOL med

frekvenser från SUC. Att bara använda listan på sammansättningar rakt av ger för hög vikt för ovanliga ord, att bara använda SUC ger inte information på ordledsnivå, och att bara använda 300 sammansättningar ger för lite data. Förmodligen skulle resultaten bli bättre om en större korpus med sammansatta ord uppdelade på rätt sätt fanns tillgänglig.

Av de 84 000 orden ur SAOL förekommer endast 10 000 i SUC, med i genomsnitt fyra förekomster per ord för dessa. Ungefär 14 % av sammansättningarna i SUC finns med i SAOL (beräknat med stickprov).

Om man som tidigare först sållar bort alla förslag till tolkningar med många ordled och använder ordledsfrekvenser endast för att välja mellan de förslag som finns kvar fungerar metoden mycket bra, 95 % rätt.

3.5 Ordklasskontext

Ofta framgår det tydligt av kontexten vilken ordklass en sammansättning bör ha. Om olika tolkningsförslag ger olika ordklass på sammansättningen kan man använda denna kontextinformation för att välja mellan dem.

Vi prövade denna metod genom att använda den statistiska ordklassgissaren TnT (Brants 2000), som gissar ordklasserna för alla ord i en text som man matar in. Programmet använder statistik över hur ofta ett visst ord tillhör en viss ordklass och statistik över hur vanliga olika ordklassföljder är.

Vi körde TnT på texten, med sammansättningen utbytt mot sitt efterled, eftersom efterledet bestämmer ordklassen, för att få ordklassen på samtliga förslag. Vi körde också TnT på texten med sammansättningen utbytt mot ett låtsasord, för att få TnT att bestämma vilken ordklass som vore mest sannolik, givet att det stod ett helt okänt ord i stället för sammansättningen. Alla förslag som inte hade denna ordklass kasserades sedan och den första enkla metoden i avsnitt 3.1 användes för att välja mellan de förslag som återstod.

Denna metod fungerade inte bra, endast 87 % rätt. I de allra flesta fallen tillhör nämligen samtliga tolkningsförslag samma ordklass, så den är sällan tillämpbar. I de fall där olika förslag fick olika ordklass från TnT berodde detta ofta på att TnT förvirrats och helt enkelt gissat fel på något av förslagen. Vanligt var också att förslag som skilde sig från andra med avseende på ordklass hade ett väldigt kort och oftast flertydigt efterled, vilket gav möjlighet till fördelaktig tolkning ur ordklassynpunkt. Korta efterled är dock sällan den korrekta tolkningen. Om man först sållar bort förslag med många ordled fungerar det lite bättre, men fortfarande inte så bra: 90 %.

3.6 Ordledens ordklasser

Vissa kombinationer av ordklasser är vanligare än andra i sammansättningar. Till exempel står substantiv-substantiv för över 25 % av sammansättningarna, medan pronomen-pronomen är extremt ovanligt.

För att utnyttja detta behövde vi konstruera ett enkelt program som gissar ordklassen på ordled; det finns såvitt vi vet inga program för det sedan tidigare, troligen på grund av att man sällan har behov av denna typ av information. Programmet gör väsentligen bara en lexikonuppslagning för att ta reda på möjliga ordklasser, och ingen hänsyn tas till kontexten. För förled finns även ett par enkla regler för morfologisk förändring i de fall då förledet inte finns i lexikonet. Samtliga möjliga ordklasser behålls, det vill säga ingen disambiguering sker. Programmet är inte helt träffsäkert, särskilt förleden drabbas av fel, men som vi förklarar nedan är detta inte ett stort problem.

För varje förslag till tolkning av en sammansättning beräknades sannolikheten att ordklassen för förledet kombinerades med ordklassen för efterledet. Om det fanns flera möjliga ordklasser för för- eller efterledet användes den mest sannolika kombinationen. I de fall då det fanns mer än två ordled gjordes beräkningen för samtliga för- eller mellanled och efterledet, och sannolikheten för förslaget beräknades som produkten av dessa sannolikheter. Förslaget med högst sannolikhet valdes sedan som tolkning.

Sannolikheterna för kombinationer av ordklasser för för- och efterled beräknades *automatiskt* genom att det ovan nämnda programmet gissade ordklasser på delarna i 300 sammansatta ord som handannoterats med sin korrekta tolkning. Eftersom programmet konsekvent gissar fel på samma sätt i dessa referensdata som det gör på ny text senare är det inte ett stort problem att det ganska ofta gissar fel. Dock är det rimligt att förmoda att metoden skulle fungera bättre om ordklassgissningen var bättre. Likaså vore det förmodligen mycket bättre med avsevärt fler referensexempel. Denna metod fungerar bra, med 94 % rätt.

3.7 Bokstavsföljder

Vissa bokstavsföljder förekommer inte i svenska ord som inte är sammansatta, till exempel »kk» och »stp». Denna egenskap har tidigare använts för att dela upp sammansatta ord (till exempel Kokkinakis & Johansson Kokkinakis 1999). Dock innehåller de flesta sammansatta ord inte sådana bokstavsföljder.

Liknande information kan också användas till att välja mellan olika uppdel-

ningsförslag. Även om de bokstavsföljder som sträcker sig över ordledsgränserna kan förekomma i ord som inte är sammansatta någon gång kan man ofta notera att de är ovanliga i andra ord.

För att utnyttja detta samlades statistik från förled och efterled över alla fyrbokstavsföljder, så kallade fyrgram, som inte sträckte sig över en ordledsgräns. Detta beräknades från de 84 000 sammansättningarna i SAOL, viktade med frekvenser från SUC. För varje tolkningsförslag för en sammansättning beräknades sedan summan av frekvensen för alla fyrbokstavsföljder som sträckte sig över en föreslagen ordledsgräns. Det förslag som fick lägst summa valdes sedan som tolkning, eftersom dess ordledsgränser alltså innehöll de bokstavföljder som var minst sannolika att förekomma där det inte skulle vara en ordledsgräns.

Vi exemplifierar hur detta går till med ordet »genomarbetat», som får två möjliga tolkningar i den första analysen, »gen-omarbetat» och »genom-arbetat». I det första förslaget summeras korpusfrekvenserna för följande bokstavsföljder: »geno (339)», »enom (342)» och »noma (4)», summa 685. För det andra förslaget: »noma (4)», »omar (4)» och »marb (14)», summa 22. Metoden väljer alltså tolkningen »genom-arbetat», eftersom den innehåller de i icke sammansatta ord förhållandevis ovanliga »marb» och »omar» vid uppdelningsgränsen, i stället för de vanliga »enom» och »geno».

Av de 18 959 fyrbokstavsföljder som finns i sammansättningarna i SAOL förekommer 10 211 minst fem gånger i ordled i SAOL. Om man räknar förekomster av dessa ordled i SUC i stället är det 3 680 som förekommer mer än fem gånger, och endast 7 634 bokstavsfyrgram från ordleden i SAOL förekommer över huvud taget i SUC. Detta beror på att som tidigare nämnts bara en liten del av sammansättningarna i SAOL förekommer i SUC (och vice versa).

Att använda fyrgram av bokstäver fungerade bra, med 95 % rätt. Det går förstås att använda kortare eller längre bokstavsföljder också, eller kombinationer av olika längder, men vi har bara undersökt fyrbokstavsföljder.

Sammansättningar med utländska lånord och namn kan innehålla bokstavsföljder som inte finns eller är väldigt ovanliga i andra svenska ord och detta kan leda till att metoden väljer fel. I praktiken är det ett mycket litet problem, då både lånord och namn finns i det material statistiken hämtats från. Det ligger också i sakens natur att sammansättningar med ovanliga lånord och namn är ovanliga, och att tvetydigheter mellan flera olika tolkningar uppstår då dessa analyseras är därför väldigt sällsynt.

3.8 Ad hoc-regler

För att ta hand om en del svårhanterade fall, som många av metoderna ovan hade problem med, konstruerades även ett par ad hoc-regler. En regel tar hand om vanliga avlednings- och böjningssuffix som även kan vara efterled i sammansättningar, till exempel »flick-or» och »spring-ande». Regeln listar helt enkelt några vanliga suffix och förslag med dessa kasseras. Även om det säkert kan finnas tillfällen då dessa används i sammansättningar, som »glödhet (glöd-het)» är det tämligen få gånger jämfört med hur ofta de är avlednings- eller böjningsändelser, som »godhet», men de föreslås ofta som en tolkning.

Ett annat problem gäller sammanskrivningen av tre konsonanter till två, som i »toppolitiker (topp-politiker)». Detta leder ibland till tvetydighetsproblem, oftast med en mycket rimligare tolkning som »vin-nyheter» kontra »vinn-nyheter», men även två rimliga tolkningar kan förekomma, som för tidigare nämnda »glasskål». De flesta metoder har svårt att hantera denna typ av tvetydighet; de har till exempel alltid ett förslag med mycket ovanligare bokstavsföljd (tre identiska konsonanter). Eftersom många metoder inte hanterar denna typ av sammansättningar på ett vettigt sätt konstruerades en regel att alltid välja tvåkonsonantstolkningen, som är något vanligare än trekonsonantstolkningen. För vissa metoder, till exempel kontextmetoderna, behövs inte denna regel.

3.9 Kombinationsmetoder

Eftersom de olika metoderna gör fel på olika sätt kan man kombinera dem och på så sätt låta dem rätta varandra. Nästan alla sammansättningar har rätt förslag valt från någon metod. Ett enkelt sätt att kombinera metoderna är att låta dem rösta om vilken tolkning som ska väljas.

En annan något mer komplicerad kombinationsmetod undersöktes. Första steget i metoden är att förkasta alla förslag utom dem med minst antal ordled, som i avsnitt 3.1. Därefter används metoden baserad på ordledens frekvenser i avsnitt 3.4 tillsammans med metoden baserad på ordledens ordklasser i avsnitt 3.6. Dessa två metoder fick avgöra hur mycket de trodde på varje förslag och sedan vägdes detta ihop, med något större vikt för frekvensmetodens uppskattningar.

Denna kombinationsmetod har rätt i nästan 98 % av fallen för de tvetydiga orden, vilket ger något mer än 98 % rätt om man ser till alla sammansättningar i testtexten. Det är märkbart bättre än de två metoderna var för sig.

Tabell 3. Korrekt analyserade sammansättningar, av dem med minst två förslag.

Metod	Rätt (%)	Fel (%)
Antal ordled	93	7
Ord i kontexten	75	25
Ord i kontexten, få ordled	94	6
Ord i kontexten, RI	51	49
Ord i kontexten, RI, få ordled	96	4
RI mellan ordleden	46	54
RI mellan ordleden, få ordled	96	4
Ordledsfrekvenser	93	7
Ordledsfrekvenser, få ordled	95	5
Ordklasskontext	87	13
Ordklasskontext, få ordled	90	10
Ordledens ordklasser	94	6
Bokstavsföljder	95	5
Kombinationsmetod	98	2

4. Feltyper

De fel som de automatiska metoderna gör kan delas in i fyra grupper. Procentsiffror nedan gäller den bästa metoden, kombinationsmetoden i avsnitt 3.9, men är att betrakta som ganska osäkra, då denna metod endast gjorde 27 fel i våra testdata. Andra metoder gör samma sorters fel, men i lite andra proportioner.

Den första feltypen är att en sammansättning delats upp vid alla korrekta ordledsgränser, och även på minst ett annat ställe. Ett exempel är »Vig-gen-plan», där »Viggen-plan» vore det önskade. Denna typ av fel orsakas vanligtvis av att något av ordleden saknas i uppdelarens lexikon. I annat fall skulle det korrekta fallet nästan alltid väljas, eftersom de flesta metoder starkt föredrar förslag med så få ordled som möjligt. Ungefär 20 % av felen var av denna typ för kombinationsmetoden.

Feltyp nummer två är att som ovan dela upp vid korrekta ordledsgränser, men lämna några sammansatta ordled kvar utan att dela upp dem, till exempel »fotboll-s-lag» i stället för »fot-boll-s-lag». Detta är den vanligaste feltypen, strax över 40 % av felen, och kan till viss del åtgärdas genom att ouppdelade sammansättningar tas bort från ordlistorna.

Beroende på syftet kan det variera hur mycket man vill dela upp orden, så denna feltyp är inte alltid oönskat beteende. Då man vill hitta möjliga avstavningsgränser vill man förmodligen dela upp så mycket som möjligt, så att man

har många avstavningsmöjligheter. Vill man däremot använda uppdelningen vid automatisk översättning är det förmodligen bättre att behålla så stora delar som möjligt, så länge delarna är ord man känner till översättningen för. Hur mycket sammansättningarna ska delas upp kan man alltså styra genom att fylla sina lexikon med olika typer av ordled.

Den tredje feltypen uppkommer vid hanteringen av tre identiska konsonanter i rad vid en ordledsgräns. Detta är ett ganska svårt problem, och många av dessa fall blir fel. Eftersom denna typ av sammansättningar dock inte är så vanlig var det under 20 % av felen som berodde på fel val av två- eller trekonsonantstolkningar.

Slutligen händer det att programmet helt enkelt delar upp ord vid fel ställen. Ungefär 20 % av felen var av denna typ.

5. Avslutande kommentarer

Automatisk uppdelning av sammansättningar kan utföras med bra resultat för svenska, 99 % av sammansättningarna i utvärderingstexten delades upp på något sätt. Av dessa delades 98 % upp korrekt av kombinationsmetoden, och en stor del av de kvarvarande felen är av en typ som i många tillämpningar inte är speciellt allvarlig. Sett över alla ord i texten så gör sammansättningsuppdelaren bara fel på 0,1 %.

De metoder som här beskrivits är alla baserade på olika typer av statistik. Det innebär att man inte explicit behöver skriva regler för hur olika ord ska hanteras. Det innebär också att det bör vara möjligt att utan nämnvärd anpassning använda samma metoder för andra språk, och en del av metoderna har redan använts för bland annat tyska, norska och koreanska.

Flera av metoderna bygger dock på att det finns text med viss annotering att samla statistik från. Med de ganska blygsamma resurser som här användes uppnåddes bra resultat, men med större resurser skulle man förmodligen komma ännu längre.

Statistik räcker alltså långt när det gäller att dela upp svenska sammansättningar. Det finns svåra sammansättningar som analyseras fel, men nästan alla sammansättningar analyseras korrekt.

Vissa sammansättningar kräver att man tittar på den kontext de förekommer i för att man i det aktuella fallet ska kunna avgöra vilken den korrekta tolkningen är. Sådana sammansättningar är dock väldigt ovanliga. Det är heller

inte vanligt att automatiska metoder använder information från kontexten. Av de två metoder vi undersökt som gör det fungerade den ena bra och den andra mindre bra.

En mycket kraftfull hjälpmetod är att förkasta alla förslag utom dem med minst antal ordled, vilket nästan alltid är korrekt. Dock finns det en del ord som har två rimliga tolkningar med olika antal ordled, som »matris» och »mat-ris» eller »finskor» och »fin-skor». Med denna metod kommer man alltså konsekvent att göra fel på den ena av de möjliga tolkningarna. För denna typ av ord måste man också utnyttja information i kontexten för att kunna göra en korrekt analys. De metoder som gjorde detta fungerade dåligt om man inte förkastar förslag med många ordled.

Litteratur

Brants, Thorsten, 2000: TnT – a statistical part-of-speech tagger. I: 6th Applied NLP Conference. Seattle. S. 224–231.

Dalianis, Hercules, 2005: Improving search engine retrieval using a compound splitter for Swedish. I: 15e nordiska konferensen för datorlingvistik (Nodalida 2005). Joensuu. http://phon.joensuu.fi/nodalida/abstracts/03.shtml

Domeij, Rickard, Hollman, Joachim & Kann, Viggo, 1994: Detection of spelling errors in Swedish not using a word list en clair. I: Journal of Quantitative Linguistics 1. S. 195–201.

Dura, Elzbieta, 1998: Parsing Words. (Doktorsavhandling.) Institutionen för svenska språket, Göteborgs Universitet.

Ejerhed, Eva m.fl., 1992: The linguistic annotation system of the Stockholm-Umeå Corpus project. (Rapport DGL-UUM-R-33.) Institutionen för lingvistik, Umeå universitet.

Johannessen, Janne Bondi & Hauglin, Helge, 1996: An automatic analysis of Norwegian compounds. I: 16th Scandinavian Conference of Linguistics. Åbo. S. 209–220.

Karlsson, Fred, 1992: SWETWOL: A comprehensive morphological analyser for Swedish. I: Nordic Journal of Linguistics 15(1). S. 1–45.

Koehn, Philipp & Knight, Kevin, 2003: Empirical methods for compound splitting. I: 10th Conference of the European Chapter of the Association for Computational Linguistics. Budapest.

Kokkinakis, Dimitrios & Johansson Kokkinakis, Sofie, 1999: I: Sense-tagging at the cycle-level using GLDB. (Rapport GU-ISS-99-4.) Institutionen för svenska språket, Göteborgs universitet.

Sahlgren, Magnus, 2005: An introduction to random indexing. I: Methods and Applications of Semantic Indexing Workshop vid 7th International Conference on Terminology and Knowledge Engineering. Köpenhamn. http://eprints.sics.se/221/

Sjöbergh, Jonas & Kann, Viggo, 2004: Finding the correct interpretation of Swedish compounds – a statistical approach. I: 4th International Conference on Language Resources and Evaluation. Lissabon. S. 899–902.

Språkriktighetsboken. (Skrifter utgivna av Svenska språknämnden 93.) Stockholm 2005.

Yoon, Juntae, 2000: Compound noun segmentation based on lexical data extracted from corpus. I: 6th Applied Natural Language Processing Conference. Seattle. S. 196–203.