Datorn som språkgranskare

OLA KNUTSSON

Datorerna kan inte bara rätta stavfel utan också särskrivningar, böjningsfel, syftningsfel och grammatiska fel. Men fortfarande förslår språkgranskningsprogrammen inte särskilt långt. Ola Knutsson, forskare i språkteknologi vid Tekniska högskolan i Stockholm, analyserar programmens kapacitet. Han demonstrerar också hur de kan utvärderas.

Datorn har utvecklats till vårt viktigaste skrivverktyg. Nästan all textproduktion sker numera med datorns hjälp. Datorn erbjuder kraftfulla verktyg både för den som vill ändra sin text på enstaka punkter och den som vill göra genomgående ändringar. Med sök- och ersättningsfunktionen kan man ändra till exempel förekomsterna av ordet *mannen* till *männen* inom meningen eller i hela texten.

Men ersättningsmanövern kan också leda till att det uppstår fel – antingen enstaka fel som *den glada männen* eller mer genomgående och satsöverskridande fel där bruket av pronomen som syftar på *mannen*/*männen* inte blir korrekt, i det här fallet att *han*/*honom* inte automatiskt ändras till *de*/*dem*. Datorns redigeringsverktyg kan på detta sätt skapa fel med en svårupptäckt spridning i texten. De inbyggda språkgranskare som följer med ett ordbehandlingsprogram granskar texten bara ytterst lokalt. Stavningskontrollen kan endast upptäcka isolerade stavfel, och om det finns någon form av automatisk grammatisk kontroll upptäcker den ofta endast fel i fraser eller i bästa fall inom satsen.

Datorn är alltså än så länge en begränsad språkgranskare med kraftiga skygglappar. Det ska jag visa i denna artikel. Samtidigt bör det poängteras att datorn som språkgranskare är effektiv och tillgänglig dygnet runt – dessutom förhållandevis billig. Datorn är totalt känslokall

och missar inte ett fel för att texten saknar innehåll eller för att innehållet är starkt fängslande. Människor förstår i motsats till datorn vad texten handlar om. Därför kan de ibland släppa igenom några fel; totalförståelsen får dem att förbise problem i textens yta.

Språkgranskningsprogram

Forskning och utveckling av automatisk språkgranskning har bedrivits under många år, framförallt för engelska (se t.ex. Jensen et al. 1983). För mindre språk, som svenska, har dock utvecklingen legat efter. Det beror helt enkelt på att de grundläggande analysverktyg som behövs, till exempel för svensk morfologi och ytsyntax, inte blivit möjliga att tillämpa praktiskt förrän under 1990-talet. Under senare delen av 90-talet har dock minst fyra forskargrupper arbetat med utvecklingen av språkgranskningsverktyg för svenska:

De inbyggda språkgranskare som följer med ett ordbehandlingsprogram granskar texten bara ytterst lokalt.

1. Språkgranskningsprogrammet Granska utvecklas av en forskargrupp vid Kungliga Tekniska högskolan i Stockholm.
2. Det finska språkteknikföretaget Lingsoft lanserade hösten 1998 den kommersiella grammatikkontrollen Grammatifix (Arppe 2000; Birn 2000).
3. Institutionen för lingvistik vid Uppsala universitet utvecklar i samarbete med danska och norska grupper, inom ett EU-projekt, språkgranskningsprogrammet Scarrie (Sågvall-Hein 1998).
4. Institutionen för lingvistik vid Göteborgs universitet utvecklar metoder för att hitta olika typer av fel i svensk text (Andersson, Cooper & Sofkova Hashemi 1999).

Det är intressant och glädjande att de olika grupperna utnyttjar delvis olika metoder för att angripa problemen inom automatisk språkgranskning. Granska använder en grundläggande analys som bygger på statistik, medan Grammatifix analys utgår från grammatiska regler. Scarrie utför en mer klassisk satslösning och försöker analysera hela meningar om grammatiken i programmet tillåter det. Gemensamt för alla fyra grupperna är dock att programmen ofta får utnyttja en ofullständig analys av meningen för att komma åt felen. Man använder i stor omfattning specifika felregler som söker efter specifika grammatiska fel. För att visa ungefär hur långt man kan komma i dag ska jag mer i detalj diskutera ett av programmen, Granska, som jag själv arbetar med att utveckla.

Vanliga feltyper

Det som kallas inkongruens i nominalfraser, t.ex. *ett villa* eller *en litet bil* är kanske den mest eftersökta feltypen av Granska och andra program för språkgranskning av svensk text. En anledning är att feltypen i många fall går att hitta inom ramen för en kort fras, vilket gör att den kan

Tabell 1. Ett urval av de feltyper som Granska, med varierande resultat, kan upptäcka, diagnostisera och korrigera.

Feltyp	Exempel
Felaktiga sammansättningar	***Undertiden*** *som det pågick kände han ingen smärta.*
Särskrivna sammansättningar	*Han orkade inte flytta* ***sten bumlingarna****.*
Objektsform efter preposition	*Han skickade en rad vänliga brev till* ***de****.*
Inkongruens i nominalfraser	*Han såg* ***de gröna villor*** *längs ån.*
Inkongruens i predikativ	***Äktenskapet är baserad*** *på kärlek mellan man och kvinna*
Böjningsfel vid främmande ord	*Han köpte* ***ett data*** *på rea.*
Tautologi	*Han anade* ***orsaken till anledningen.***
Böjningsfel i verbkedjan	*Vi* ***har spela*** *en låt av Aphex Twin.*
Ordföljdsfel	*Han sa att* ***han sparkade inte*** *bollen*

ringas in med ganska hög precision. Den grundläggande analys som behövs är morfologisk, d.v.s. den gäller ordens böjning. Program för detta ändamål har funnits i några år. Den morfologiska analysen är dock i många fall flertydig och granskningsprogrammet måste försöka välja en tolkning av orden. Detta medför i vissa fall att falska alarm uppstår eller att fel inte upptäcks.

Ett exempel på ett falskt alarm är när Granska pekar ut *stora läromedelsförlag* som felaktigt i meningen *Stora läromedelsförlag förefaller avvakta medan initiativ från mindre företag tenderar att inte få så stor spridning*. Det falska alarmet uppstår eftersom *läromedelsförlag* är ett okänt ord för Granska. Programmet måste då göra en sammansättningsanalys för att kunna slå upp efterledet *förlag* i programmets lexikon. Ordet *förlag* är tvetydigt mellan singular och plural, detsamma gäller det föregående adjektivet *stora*. I det här fallet väljer Granska en singulartolkning av orden, vilket resulterar i att en felsökningsregel upptäcker en bestämdhetskonflikt mellan *stora* som står i bestämd form och *läromedelsförlag* som står i obestämd form. Felsökningsregeln vill egentligen upptäcka fel av typen *gröna bil* och *lärda man*. Regeln ska heller inte påpeka fel om substantivet är tvetydigt mellan singular och plural, men eftersom ordet läromedelsförlag är okänt *ser* reglerna i nuvarande version av Granska inte tvetydigheten. (Detta bör dock gå att göra något åt i en framtida version.) Om Granska inte angrep konstruktioner med okända ord skulle många fel inte upptäckas, eftersom vi ständigt nybildar ord när vi skriver, till exempel genom sammansättning.

De automatiska språkgranskare som finns i dag baserar sitt urval av feltyper på i huvudsak två principer. Den första principen utgår från de feltyper som faktiskt går att upptäcka med tillgängliga automatiska medel. Den andra utgår från feltypens frekvens och impopularitet hos språk-

vårdare, svensklärare och allmänhet. Den första principen är starkare än den andra. Den språktekniska forskningen har andra mål än språkvården, och som en konsekvens av detta är det delvis andra feltyper som är viktiga för den språktekniska forskningen. Språkteknologerna angriper en del feltyper för att pröva vissa metoder eller de tekniska gränserna för vilka fel som är möjliga att upptäcka.

I tabell 1 presenteras ett urval av de feltyper som Granska och andra svenska skrivkontrollprogram kan hantera. I fråga om särskrivna sammansättningar har Granska kommit längst.

Kunskapskällor

Vilken kunskap bygger ett språkgranskningsprogram på? Ja, någon grammatik för svenska språket som datorn förstår finns inte. Formuleringarna av de grammatiska reglerna i ett språkgranskningssystem är delvis baserade på en tillämpning och formalisering av regler i olika grammatikböcker. När utvecklingen av Granska startade använde vi Thorells ”Svensk grammatik” (1977) som utgångspunkt för till exempel vilka nominalfrasmönster som borde vara med och i någon mån skulle kontrolleras av programmet.

Men minst lika viktigt som grammatikböcker är att använda stora textmassor som utvecklingsmaterial. Det visade sig ganska snart att många felkonstruktioner missades och många falska alarm uppstod i den version som baserades på Thorells grammatik. Det är ingen kritik mot Thorells grammatik; den var en utmärkt startpunkt. Ett mer omfattande verk hade i själva verket dränkt utvecklingsarbetet med grammatiska regler och undantag. Under 1999 kom Svenska Akademiens grammatik (Teleman, Hellberg & Andersson 1999) och många frågetecken och falska alarm fick en ny belysning. Utvecklingen av Granska hade vid denna tidpunkt nått ganska långt och programmet var redo att testas på de många exempel som ges i Svenska Akademiens grammatik (SAG).

Detta visade sig vara en utmärkt metod för att upptäcka luckor i Granska och för att få svar på varför många oförutsägbara falska alarm uppstod. SAG fungerade också utmärkt i valsituationer där man måste avgöra om man ska acceptera ett missat fel eller ett falskt alarm för att det är ett erkänt ”problem”. Vissa problematiska konstruktioner som *Statsrådet är ensam på rummet* eller *Ärter är gott* finns väl beskrivna i SAG. Det innebär dock inte att sådana konstruktioner automatiskt kan behandlas korrekt av ett datorprogram. Glappet mellan vad som låter sig beskrivas i en grammatikbok och vad som kan formuleras så att en dator förstår är fortfarande mycket stort.

Feltäta texter

Ett kärnproblem för automatisk språkgranskning är hur text med fel i över huvud taget skall ges en grundläggande språklig analys. Ett system som skall identifiera felaktiga konstruktioner i språket måste kunna hantera ogrammatiska och oförutsägbara konstruktioner. Det är också rimligt att datorn väljer en grammatisk tolkning framför en ogrammatisk tolkning. Men många fel undgår upptäckt eftersom datorn inte kan dra in ett större textsammanhang i analysen för att avgöra om en konstruktion är grammatisk eller ogrammatisk.

För en dator kan till exempel en felaktigt skriven eller okänd förkortning göra det svårt att avgöra vad som är ett ord el-

ler en mening. Detta får konsekvenser för den vidare analysen i form av meningar som saknar till exempel huvudverb. En mening med en felaktig förkortning av *till exempel* som *Jag har sett te. x. lejon och tigrar* tolkas som tre meningar. Den första meningen består av *Jag har sett te*, den andra meningen består av *x* och den tredje meningen innehåller endast orden *lejon och tigrar*. En tänkbar konsekvens blir att programmet anmärker på att mening två och tre saknar finit verb, eftersom fel endast eftersöks inom meningen.

När datorn fortsätter att leta fel i satser och meningar stöter den genast på nya problem. En framgångsrik ansats för en dator när den letar fel är att dela upp meningen i satser. Detta förutsätter dock att meningen är något så när välformulerad, annars famlar algoritmerna för satsgränsigenkänning i mörker. Om det till exempel saknas ett finit verb eller en konjunktion i en mening eller en sats blir det svårt för datorn att peka ut någon gräns mellan satserna.

Det finns många fler exempel på hur den grundläggande analysen av en text med fel i stöter på problem som omöjliggör den fortsatta felsökningen, som i många fall bygger på att det grammatiska felet står isolerat i den annars välformulerade meningen. Det innebär att Granska kommer att upptäcka fel som särskrivningen *hus bil* i meningar där det står isolerat: *Jag bor i en hus bil* . Men programmet upptäcker inte samma fel när det sammanfaller med inkongruens i nominalfraser som i *Jag bor i ett hus bil*. I det första fallet försvinner både inkongruensen och särskrivningen om *hus bil* skrivs som ett ord. I det andra fallet finns inkongruensen i nominalfrasen kvar även om den särskrivna sammansättningen skrivs ihop till ett ord. Detta medför att Granska i det andra fallet inte ”vågar” signalera att *hus bil* är en särskriven sammansättning, eftersom särskrivningen inte är ett isolerat fel i satsen.

Testa program

Hur gör man för att testa vad ett språkkontrollsprogram klarar av? Dels kan man pröva hur många och vilka fel det hittar i en större samling texter. Dels kan man göra ett s.k. användartest och se hur olika skribenter hanterar den information som programmet ger. Även om Granska ännu inte är någon färdig produkt, med den finslipning som en kommersiell programvara kräver, har jag testat programmet på en större mängd text och gjort en liten användarstudie.

Hur mycket text måste man testa på? Vissa konstruktioner är frekventa och kräver därmed mindre textmaterial. Inkongruens i nominalfraser som består av en artikel, ett adjektiv och ett substantiv – som till exempel *det glad barnet* – kommer förmodligen att testas tillräckligt, likaså särskrivna sammansättningar. Andra feltyper är ovanligare och kräver därför mycket text, till exempel inkongruens i predikativa attribut: *Vi upptäckte kistor fylld med guld*.

Att finna representativa texter för den kontext som ett granskningsverktyg skall användas i är ingen enkel uppgift, varför jag fick använda principen ”man tager vad man finner”. Jag valde ut texter på sammanlagt ungefär 200 000 ord (cirka 400 A4-sidor). De representerar fem olika texttyper: sportnyheter 63 568 ord, utrikesnyheter 20 881 ord, myndighetstexter 36 667 ord, populärvetenskap 32 386 ord, gymnasie- och högskoleuppsatser 47 517 ord. Efter att ha gått igenom texterna ma-

Tabell 2. En överblick över de feltyper som fanns med i utvärderingstexterna.

Feltyp	Exempel	Antal fel	Andel av alla fel	Eftersöks av Granska
Böjningsfel i verbkedjan	*Hon har spelar fotboll.*	89	21 %	Ja
Särskrivna sammansättningar	*Kokt kyckling lever.*	74	18 %	Ja
Inkongruens i nominalfraser	*Det här är en falska sats.*	69	17 %	Ja
Ord saknas	*Han kör__ grön bil*	56	13 %	Ja, begränsat
Stavfel med grammatisk-semantisk konsekvens	*Lisa gillar Pelle, hon vill träffa hon nu.*	55	13 %	Nej
Inkongruens i predikativ	*Denna sats är falska.*	16	4 %	Ja
Felaktig pronomenform	*Han gav boken till de.*	14	3 %	Ja, begränsat
Versalfel	*… meningen slut. den*	11	3 %	Nej
Felaktig preposition	*Han var imponerad med mig.*	11	3 %	Ja, begränsat
Ordföljdsfel	*Han sa att han sparkade inte bollen.*	8	2 %	Ja, begränsat
Överflödigt ord	*Han kör kör en grön bil.*	4	1 %	Nej
Bestämdhetsfel	*Här kommer taxibil.*	4	1 %	Nej
Kontamination, sammanblandning av uttryck	*Thurston Moore tillhör en av samtiden största.* (ska vara *är en av* eller bara *tillhör*)	4	1 %	Ja, begränsat
Kommateringsfel	*Han gillar bilar prylar och mat.*	2	0,5 %	Nej
Syftningsfel	*Bilen körde på vägen, det körde sen av.*	1	0,25 %	Nej
Totalt		418		

nuellt skilde jag ut femton feltyper, grovt sett. De redovisas i tabell 2.

Utvärderingen visade på ganska stora skillnader i felförekomst mellan olika texttyper. Man bör dock ha i åtanke att materialet är begränsat och obalanserat och att individuella skillnader mellan skribenterna påverkar resultatet i hög grad. Sammanfattningsvis kan man säga att Granska upptäcker de flesta felen i texter med få fel i, medan många fel undgår upptäckt i mer feltäta texter. Antalet falska alarm ökar i mer felfri text, medan de blir färre i texter med många fel i. I de populärvetenskapliga texterna upptäckte Granska nästan nio av tio fel, och ungefär hälften av felrapporterna var korrekta. I gymnasie- och högskoletexterna var situationen den omvända, endast fyra av tio fel upptäcktes, men sju av tio felrapporter var korrekta.

Det finns flera tänkbara förklaringar till varför detta resultat uppstår. En första förklaring är att det är svårt att upptäcka fel där också kontexten innehåller fel. Den andra förklaringen är att det är enklare att förutsäga slarvfel än rena kunskapsfel. Gymnasieelever och högskolestudenter åstadkommer fel som i många fall måste förklaras som kunskapsfel eller bristande granskningsförmåga. Dessa fel kan vara till exempel särskrivna sammansättningar, stavfel med grammatisk-semantiska konsekvenser eller helt enkelt frånvaro av nödvändiga ord. Felen i populärvetenskapliga texter verkar mer ha karaktären av slarvfel, som till exempel inkongruens i *den ansträngd diskussionen* och böjningsfel i verbkedjan *Jag har undersök en sak.* Denna andra förklaring är ännu så länge en spekulation; man måste närmare undersöka hur skribenterna arbetar och resonerar under granskningsprocessen.

I fortsättningen kanske vi inte ska försöka utveckla ett enda språkgranskningsprogram. I stället vill vi pröva olika versioner.

Olika användare

I en mindre användarstudie undersökte jag Granska och Grammatifix med fem användare (Knutsson 2001). Användarna skulle ta ställning till programmens detektioner (upptäckter), diagnoser (beskrivning av felet) och ersättningsförslag. Båda systemen gav en del falska alarm, varav några missledde användarna. De falska alarm som härstammade från stavningskontrollerna accepterades inte av användarna; däremot godkände några av användarna andra falska alarm från både Granska och Grammatifix. Dessa falska alarm gällde till exempel inkongruens i nominalfraser och särskrivna sammansättningar.

I Granska ges i vissa fall flera diagnoser till ett fel, och det verkade som om ersättningsförslagen är viktiga för att användarna ska kunna avgöra vilken av diagnoserna som är korrekt. Det fanns också tendenser som pekar mot att språksäkra användare bara behöver en detektion för att korrigera felen; de kräver inte diagnoser eller ersättningsförslag. Detta visade sig genom att användarna själva gick in

direkt i texten i stället för att använda de ersättningsförslag som programmen gav.

De tendenser som kom fram i textutvärderingen och användarstudien visar att det är intressant att studera olika användargrupper och anpassa Granska efter dem. Olika skribenter har nämligen mycket olika behov. I fortsättningen kanske vi inte ska försöka utveckla ett enda språkgranskningsprogram. I stället vill vi pröva olika versioner för grundskoleelever, yrkesskribenter, andraspråksinlärare och många andra grupper. Detta kommer dock att ställa nya och högre krav på användargränssnitt, granskningsmetoder, hjälpsystem och språkligt innehåll. ■

Litteratur

Andersson, R., Cooper, R. & Sofkova Hashemi, S. (1999). Report: a system for finding ungrammatical noun phrases in Swedish text. Department of Linguistics, Göteborgs universitet, Göteborg.

Arppe, A. (2000). Developing a grammar checker for Swedish. I *Proc. 12th Nordic Conference in Computational Linguistics, Nodalida 1999*. Trondheim, s. 13–27.

Birn, J. (2000). Detecting grammar errors with Lingsoft's Swedish grammar checker. I *Proc. 12th Nordic Conference in Computational Linguistics, Nodalida 1999*. Trondheim, s. 28–40.

Domeij, R., Knutsson, O., Carlberger, J. & Kann, V. (2000). Granska – an efficient hybrid system for Swedish grammar checking. I *Proc. 12th Nordic Conference in Computational Linguistics, Nodalida 1999*. Trondheim, s. 49–56.

Jensen, K. Heidorn, G.E., Miller, L.A. & Ravin, Y. (1983). Parse fitting and prose fixing: Getting hold on ill-formedness. *American Journal of Computational Linguistics 9*, no. 3–4, s. 123–136.

Knutsson, O. (2001) Automatisk språkgranskning av svensk text. Licentiatavhandling, Institutionen för numerisk analys och datalogi, Kungliga Tekniska Högskolan, Stockholm.

Sågvall Hein, A. (1998). A chart-based framework for grammar checking. Initial Studies. I *Proc. 11th Nordic Conference in Computational Linguistics, Nodalida 199*8, Köpenhamn, s.68–80.

Teleman, U., Hellberg, S. & Andersson E. (1999). *Svenska Akademiens grammatik.* Band 1–4, Nordstedts Ordbok, Stockholm.

Thorell, O. (1977). *Svensk grammatik.* Andra upplagan, Stockholm.

Det som presenteras i denna artikel vilar på mångårig forskning kring skrivande och språkliga datorstöd hos en forskargrupp ledd av Kerstin Severinson Eklundh på Kungliga Tekniska högskolan i Stockholm. Gruppen har under åren utvecklat ett nära samarbete med en forskargrupp inom teoretisk datalogi med inriktning mot effektiva algoritmer för språklig analys, ledd av Viggo Kann. Grupperna utvecklar tillsammans språkgranskningsprogrammet Granska (Domeij, Knutsson, Carlberger & Kann, 2000).