ATT TALA MED MASKINERNA

Peter Gärdenfors

Kognitionsforskning
Kungshuset, Lundagård
222 22 LUND

E–mail: peter.gardenfors@fil.lu.se

Samtal i köket

Det var en morgon alldeles i början av 2000–talet. När jag steg in i köket tändes taklampan. Den fick mina mosiga ögon att knipa i hop sig som musslor. Disken stod kvar sen i går kväll. Kylskåpet spann trånsjukt. Jag vände mig mot spisen och sa:

–Spisen.

–Redo, svarade den.

–Sätt på lilla plattan på 250 grader.

–Ska ske, sa spisen med klanglöst tonfall och en röd lampa tändes vid lilla plattan.

Jag öppnade kylskåpet för att ta ut mjölken. Från skåpets innanmäte började en metallisk röst att mässa:

–Det finns ett mjölkpaket kvar. Sista förbrukningsdatum är den 21 oktober. Bäst-före-datum för leverpastejen var i tisdags i förra veckan. Detta är sjunde gången det rapporteras. Det finns inget smörgåsmargarin. Det är slut på ...

Lund University Cognitive Studies – LUCS 53. ISSN 1101–8453.

Jag högg mjölkpaketet och stängde dörren. Rösten tystnade. Jag tog fram grovbrödet och skar upp ett par skivor. Smöret var tydligen slut, men det fanns kanske litet ost kvar.

–Lilla plattan är varm, sa spisen och den röda lampan började blinka.

Jag öppnade kylskåpet igen för att leta efter osten.

–Det finns ingen mjölk kvar, började det omedelbart inifrån.

–Tacka fan för det, sa jag. Jag har ju precis tagit ut den.

–Bäst-före-datum för lever...

Jag hade hittat en tub Kalles kaviar och smällde snabbt igen dörren.

–Lilla plattan är varm, sa spisen.

–Håll klaffen, sa jag.

–Lilla plattan är varm, sa spisen.

* * *

Tidiga stadier i kommunikation med datorer

Vi har vant oss vid att kommunicera *via* maskiner. Först telegrafen och sedan telefonen. Med mobiltelefonernas hjälp har radiokommunikation blivit allmänt tillgänglig. I Sverige kan vi samtala med nästan vem som helst från nästan var som helst. Allt fler använder datorerna som *medier* för kommunikation; vanligast är elektronisk post, men man kan också få en mer direkt kontakt via någon chat-linje. Om man deltar i en videokonferens kan man till och med se den man talar med.

Men vårt sätt att kommunicera *med* maskinerna har också utvecklats. Särskilt tydligt märks detta i hur vi umgås med datorerna. Man kan säga att människans kommunikation med datorer har gått igenom tre faser. Den första var *trollformlernas* tid. För att få maskinen att göra något över huvud taget gällde det att ge den kryptiska kommandon i ett hemligt språk. Minsta stavfel eller en felande parentes bestraffades med totalt sammanbrott i kommunikationen. Den som har använt MS–DOS eller Unix-kommandon vet precis vad jag talar om.

Den andra kommunikationsfasen, som är den som dominerar för närvarande, befinner sig på *pekboksnivån*. Macintoshens ”skrivbord” och PCns ”fönster” gör att vi interagerar med datorn på ett nytt sätt. Med musens hjälp kan man peka på bilder eller ikoner, och genom att

klicka på dem kan man få fram nya objekt att peka på. Man drar, man flyttar, man förstorar och man markerar ytor. Hela interaktionen är *rumsligt* orienterad och den använder mer naturliga handrörelser än det vanliga knackandet på tangentbordet. Vissa skärmar kan man peka eller rita på vilket gör kontakten ännu mer direkt. Men det är fortfarande en typ av kommunikation man använder med ett litet barn som inte förstår vad man säger.

Talad kommunikation

Anledningen till att vi använder en sådan primitiv interaktionsform är att datorer hitintills har varit döva och blinda. Men med talteknologins hjälp är vi nu på väg in i den tredje fasen där datorerna blir hörande – jag kallar det *militärkommandots* stadium. Vi kommer att kommunicera allt mer med datorerna med hjälp av *talat språ*k. De finns redan flera system för persondatorer som kan tolka och utföra ett antal muntliga kommandon. Man kan till exempel säga: ”Öppna brevlådan med e-post”, ”läs det första brevet” och datorn läser upp brevet med en robotröst samtidigt som man kan se det på skärmen. Det finns också program där man kan diktera ett brev direkt till datorn. Resultatet blir inte felfritt, men det går relativt lätt att putsa i ett ordbehandlingsprogram efteråt.

Systemen för taligenkänning utvecklas snabbt. Inom en snar framtid kan man i många sammanhang ersätta tangentbord och datormus med röststyrning. Ja, vem har sagt att man måste kommunicera med en dator via ett tangentbord och en bildskärm? Det räcker långt med mikrofon och hörlurar. Det som behöver ses kan projiceras på insidan av ett par glasögon eller till och med direkt på näthinnan. Och i den mån vi fortfarande använder skärmar kommer vi att peka direkt på dem utan att gå omvägen via en mus.

På ett liknande sätt kommer vi att tala med en mängd olika vardagsföremål. Du kan säga till hissen vilken våning den skall ta dig till. Ett barn kan be telefonen ringa upp mamma på jobbet. Det är heller inga principiella problem att tala med bankomaterna – men frågan är om vi *vill* göra det. Det skall vara hemligt hur mycket pengar man tar ut. Det kommer att finnas automater för turistinformation som kan förstå frågor på flera olika språk – och svara med samma mynt.

Datoröversättning av talat språk

De första systemen för automatisk översättning kom redan i slutet av 50–talet. Men de fungerade dåligt eftersom de byggde på en felaktig metod. Det fanns en period då man hade gett upp hoppet om att kunna få datorer att göra översättningar. Men nu är industrin i full gång, även om det är långt kvar till kompletta och felfria system.

Vid flera laboratorier runt om i världen arbetar man med system som skall kunna översätta direkt från tal på ett språk till tal på ett annat. En av de största satsningarna görs vid Carnegie Mellon University i USA i samarbete med Universität Karlsruhe i Tyskland där man utvecklat att system som kallas Janus II. I Tyskland pågår också ett samarbete mellan 32 olika universitetsgrupper inom ett liknande projekt som kallas Verbmobil.

Systemet Janus II kan översätta mellan engelska, tyska, spanska, japanska och koreanska. Det är specialiserat på samtal som handlar om två personer som skall avtala ett möte. För detta krävs en vokabulär på mellan tre och fem tusen ord beroende på vilket språk som används. Systemet översätter inte riktigt lika snabbt som man talar normalt utan det tar ungefär dubbelt så lång tid.

Till skillnad från de flesta äldre system översätter Janus inte bara med hjälp av en syntaktisk analys utan den talade frasen ges en *semantisk* tolkning. Frasens semantiska innehåll beskrivs på ett sätt som är gemensamt för alla de språk som systemet kan hantera. Beskrivningen är en sorts ”interlingua” som sedan i sin tur används för att generera en talad fras på det språk som ger översättningen. Eftersom syntaxen inte representeras i den semantiska tolkningen, kommer den översatta frasen ofta att ha en helt annan grammatisk struktur än den som talades in på det ursprungliga språket. Systemet är långt ifrån perfekt. Man uppskattar att man i bästa fall uppnår en innehållsmässigt korrekt översättning av ungefär 70% av de talade fraserna.

Inom Janus-projektet arbetar man med att utveckla flera olika tillämpningar av översättningssystemet. Ett exempel är en station för videokonferenser som gör översättningar medan man samtalar. Varje partner i konferensen ser de andra och hör deras röster, och får strax höra en talad översättning av vad som sagts, samtidigt som den översatta texten visas på videoskärmen. Den som har talat får till och

med möjlighet att kontrollera att den semantiska tolkningen av vad som sagts är rimlig innan översättningen sänds över till de andra. Eftersom alla deltagare i videokonferensen vill förstå varandra kan man, i de fall översättningen inte är begriplig, be talaren formulera om frasen eller i värsta fall få honom eller henne att skriva in den.

Ett annat exempel är en bärbar översättare. Det finns en version som implementerats i en bärbar Pentiumdator och som är en förenklad version av det stora Janus-programmet. Datorn kan ge en något långsam talad översättning av det som exempelvis en turistguide berättar. Man experimenterar också med att visa den översatta texten på insidan av ett par glasögon.

Slutligen försöker man använda Janus-systemet vid simultan översättning av en naturlig konversation. Denna sorts översättning anses vara svår för mänskliga tolkar eftersom man normalt gör snabba växlingar i en dialog och ofta talar i mun på varandra. I en sådan situation arbetar systemet direkt med yttrandena på de båda språken utan mänsklig inblandning. Översättningarna visas som en text på datorskärmen. Så länge dialogen handlar om ett område där systemet känner till de ord som används fungerar det någorlunda, men det lär dröja innan denna typ av program blir tillräckligt bra för att vara kommersiellt gångbara konsumentprodukter.

* * *

Tillbaka till morgonen i köket. Tevattnet var nu varmt och jag hade lagt i en påse Lipton's. Jag vände mig mot TVn och aktiverade den med det vanliga tilltalsordet:

–Dumburken!

–Vad önskar du se på? svarade den, som alltid omedveten om förolämpningen.

–Nyheter om Israel, sa jag.

Efter ett par sekunder kom det upp sju alternativ på TV-skärmen med olika nyhetsprogram under det senaste dygnet där ordet "Israel" har nämnts. TVns dator hade hittat dessa alternativ genom att gå igenom alla de digitaliserade nyhetsprogram som hämtats in via kabeln och med hjälp av taligenkänning lokaliserat de snuttar som matchade mitt önskemål.

–TV2s sena nyheter från i går kväll, sa jag och reportaget rullade igång. Det visade sig inte vara särskilt dramatiskt, utan var ett inslag om hur israelerna utvinner dricksvatten från havet med hjälp av solenergi. Jag försjönk åter i dagdrömmar över min tekopp och kaviarmackan.

* * *

Vi kommer att kunna prata med våra bilar. Bilen kommer att varna för att en dörr inte är stängd, att det finns för lite bensin i tanken, eller att en annan bil håller på att köra om. Vi kommer att kunna be den om väganvisningar som den beräknar med hjälp av en GPS-mottagare och en inlagrad karta. Bilens system kommer att leta fram en rutt som tar hänsyn till trafiktätheten på olika gator för att på så sätt kunna föreslå den snabbaste vägen. Dessutom kan vi med samma system beställa mobiltelefonsamtal helt muntligt eller styra en vanlig dator där resultaten visas på vindrutan i den mån de inte kan presenteras som tal.

Det finns redan bilar som ger talade upplysningar om bilens tillstånd. Tekniskt sett är det inga problem att göra talsystem som talar om allt möjligt om bilen. Men det verkar som om dessa system inte är särskilt populära. Anledningen är att de upplevs som *tjatiga*. De upprepar samma information om och om igen precis som spisen och kylskåpet i köket. Resultatet blir att förarna snart stänger av talsystemen.

Måste sådana system vara tjatiga? Vore det inte lätt att undvika upprepningar? Nej, faktum är att för att slippa tjatet krävs ett betydligt mer avancerat system. Det talande kylskåpet har ett budskap att förmedla om leverpastejen. För att veta att det inte behöver upprepa meddelandet nästa gång jag öppnar kylskåpet måste det kunna avgöra att det är samma person som öppnar dörren. Vore det en annan person, kunde det vara högst väsentligt att tala om en gång till att leverpastejen är kass. Men det är mycket svårt för maskiner att känna igen personer, och den förmågan kräver en helt annan teknologi som inte finns riktigt ännu.

Ett annat problem är att vissa budskap är *irrelevanta*. Det kan vara bra att ha en bil som varnar för att en dörr inte är riktigt stängd innan du börjar köra. Men om du öppnar dörren för att kliva ut, så vill du inte att bilen skall varna för att dörren är öppen. Men hur skall talsystemet veta om du *vill* ha dörren stängd eller inte? För att kunna hantera detta måste det skapa sig en bild av dig och dina önskningar. Detta kallas inom

datorvärlden för en *användarmodell*. En sådan modell är nödvändig för att kunna skapa ett talsystem som kommunicerar på någotsånär mänskliga villkor.

Dialog

Den talade interaktion vi nu har med olika maskiner klassificerade jag som militärkommandon. Men kommandon är inte någon särskilt avancerad form av kommunikation. De är för enkelriktade.

En framtida fjärde fas i samtal mellan människor och maskiner bör sträva efter att efterlikna den mänskliga *dialogen*. En riktig dialog förutsätter att den man talar med kan komma med kommentarer, invändningar och motfrågor. De talande system som finns i datorvärlden sitter där som papegojor som mekaniskt upprepar sina fraser utan att vänta på någon respons. De för inga samtal. Ordet ”samtala” betyder ju just att tala tillsammans.

Varför är det så svårt att göra dialogsystem? Det som framför allt saknas för att datorerna skall kunna samtala med oss är en *föreställning* om hur vi tänker och vad vi vill. De användarmodeller som finns är alldeles för primitiva för att räcka till i dialogsammanhang.

System för talsyntes och taligenkänning fungerar därför att det finns en fonetisk *teori* som förser programmeraren med lämpliga variabler för ljudanalysen. Utan en djupgående kunskap om de mekanismer som ligger bakom hur människor producerar och uppfattar talat språk hade det inte funnits någon möjlighet att skapa datorprogram som löser motsvarande uppgifter. Vi ser nu frukterna av det teoretiska arbetet inom fonetiken genom att man kan konstruera hyfsade program för talsyntes och taligenkänning.

Det finns tyvärr ingen motsvarande teori för hur dialoger fungerar. Ett steg i rätt riktning är den analys av språklig mening som lagts fram av filosofen Paul Grice. Hans teori betonar att *samspel* och *återkoppling* är nödvändiga för att meningsfull kommunikation skall uppstå. Om jag skall lyckas meddela min vän något, räcker det inte med att jag säger det. Jag måste också veta att hon förstår vad jag yttrar och att hon tror att jag talar sanning och inte driver med henne. Vi får komplicerade mönster av typen “Jag *tror* att hon *tror* att jag *menar* det jag säger”. Sådana nästlade föreställningar är karakteristiska för mänsklig information.

Filosoferna kallar sådana föreställningar för ”högre ordningens intentioner”. Det som är väsentligt att komma åt är *avsikten* bakom det kommunicerade budskapet. Den språkliga form som budskapet ges är bara ett redskap för detta.

Tänk bara på hur *ironi* fungerar. Här måste den som talar kunna tänka “Jag tror att hon *inte* tror att jag menar vad jag säger” för att man skall våga vara ironisk. Denna föreställning är ett exempel på en tredje ordningens intention. Enligt Grice är sådana föreställningar nödvändiga för att vi skall kunna samtala och inte bara bete oss som papegojor. Samtal bygger på att vi kan spegla oss i varandras tankar.

En annan faktor som spelar en central roll i dialoger är *förväntningar* . Om jag yttrar “Kan du skicka såsen?” till min bordsdam, tror jag inte att hon tror att jag menar vad jag säger (jag skulle bli förvånad eller irriterad om hon bara svarade “Ja”), utan jag *tror* att hon *tror* att jag genom att fråga om hon kan skicka såsen *förväntar* mig att hon skall *förstå* att jag *vill* att hon skall skicka den. Denna förväntning, som är av en helt vardaglig typ, innehåller faktiskt fem nivåer av speglingar av föreställningar om den andre!

Försök få din dator att hänga med på ett sådant spegelspel! En dator tror sällan något, än mindre tror den något om vad du tror. Så det lär dröja innan vi får program som kan förstå ironiska kommentarer.

För att ha en chans att skapa datorprogram som kan föra ett samtal är det inte tekniken som är den felande länken utan kunskap om vad som händer när människor förstår varandra. Vi behöver framför allt forskning inom semantik och pragmatik för att öka vår teoretiska förståelser för dialogens mekanismer.

Bilden av den andre

En dialog förutsätter att de som samtalar har en *gemensam värld*. Naturligtvis måste man tala samma språk, men ännu viktigare är att man har gemensamma föreställningar om hur världen fungerar. Om världarna inte överlappar, får man ägna mycket tid åt att förklara sånt som man själv tycker är självklart. Och omvänt, ju mer gemensam värld man har, desto mindre behöver sägas. Personer som umgås dagligen behöver inte säga mycket för att kommunicera.

Men ett samtal utgår inte bara från en gemensam värld. Det *bygger upp* en också. Under konversationens gång för man in personer och ting, som man sedan kan referera tillbaka till. I språket är det framför allt pronomen som har en sådan refererande funktion. När man väl har beskrivit en okänd kvinna för den man talar med, kan man sedan använda ord som ”hon” och ”den där du vet” i stället för den långa beskrivningen. Pronomen gör språket mycket mer ekonomiskt.

Men datorer har det svårt med pronomen. Antag att datorn får höra följande:

”Bonden var tvungen att hämta veterinären till tjuren. Han hade skadat pungen på ett taggtrådsstängsel”.

En mänsklig läsare har inga större problem med att förstå att ”han” i den andra satsen syftar på tjuren. Ingen grammatik kan avgöra detta utan det fordras att man *förstår* texten; språkets semantik är en förutsättning för att ”han” skall översättas rätt. Det krävs *kunskap* om hur människor och tjurar lever för att förstå att det inte är bonden eller veterinären som har skadat pungen. Förståelsen ges av en modell av den värld som man talar om. När vi läser en text skapar vi oss en bild eller föreställning av det vi läser om. En sådan föreställning utgör en väsentlig del av den gemensamma värld som ett samtal bygger på.

Det som i datorvärlden kallas användarmodell är första steget mot den gemensamma värld som krävs för att en dialog skall kunna uppstå. Den gemensamma världen byggs upp genom att lyssnaren skapar sig en bild av talarens föreställningar och vice versa. Allt eftersom samtalet fortskrider anpassar man sina föreställningar till varandra. Den gemensamma världen blir allt mer omfattande. Som en följd kommer dialogen att blir mer fåordig och mer obegriplig för en utomstående. Det självklara behöver inte sägas.

En aspekt av att delta i en dialog är att båda partner kan *ta initiativ*. I ett kommandosystem är det bara ena sidan som är aktiv – den andra tar order och utför dem efter bästa förmåga.

Mitt budskap är att för att vi skall kunna tala med datorerna på samma sätt som vi samtalar med människor krävs framför allt att datorsystemen får möjlighet att skapa modeller av hur den mänskliga samtalspartnern föreställer sig världen. Man kan inte förvänta sig att användaren spontant talar om allt som han eller hon utgår från i samtalet.

Datorprogrammet måste därför vara aktivt och ställa frågor för att få reda på de förutsättningar som användaren arbetar med.

Men vi vill naturligtvis inte att datorn tar kommandot. Lik en engelsk butler vill vi att den skall vara förstående, tolerant och hjälpsam. Den skall enligt robotlagarna ha vårt bästa som sitt främsta mål.

Om vi lyckas ge datorer en modell av våra inre världar, kommer de att bli mycket ”mänskligare”. Vi kan exempelvis förvänta oss ordbehandlingsprogram som ger smarta stilistiska och innehållsmässiga råd angående den text man håller på att skriva (”Det där skrev du ju redan för tre sidor sedan. Måste du jämt upprepa dig?”). Men vi kommer inte att få något som går upp mot skickligheten hos en gammaldags sekreterare förrän datorn *förstår* den skrivna texten.

Vägen till datorsystem som har tillräckligt rika användarmodeller är mycket lång, och det krävs mycket forskning för att så pass väl förstå hur en mänsklig dialog fungerar att man kan få en dator att bete sig på ett liknande sätt. Men vi kan kanske hoppas på att man kan konstruera fylliga användarmodeller inom speciella områden. Först då kommer vi att lämna kommandofasen och få något som närmar sig en dialog med maskinerna.

* * *

–Vädret i Skåne, sa jag till TVn. Den växlade program och visade det som jag kunde konstatera genom att se ut genom köksfönstret.

–Stäng av, sa jag medan jag torkade brödsmulorna från bordet, och skärmen slocknade med en suck av statisk elektricitet.

–Lilla plattan är varm, sa spisen.

–Håll klaffen, sa jag.

Litteratur:

P. Grice: ”Utterer's meaning and intentions”, *The Philosophical Review* 78 (1969), ss. 147–177.

P. Gärdenfors: *Fängslande information*, Natur & Kultur 1996.

A. Waibel: ”Interactive translation of conversational speech”, *Computer*, July 1996, ss. 41–48.

S. Winter: ”Det självklara behöver inte sägas”, *Lund University Cognitive Studies* 46, 1996.