NyS

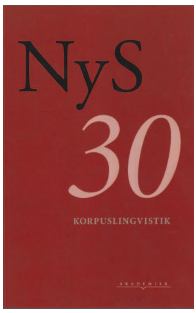
Titel: Göteborgskorpusen för talspråk

Forfatter: Jens Allwood, Leif Grönqvist, Elisabeth Ahlsén og Magnus Gunnarsson

Kilde: *NyS – Nydanske Sprogstudier* 30. *Korpuslingvistik,* 2002, s. 39-58

Udgivet af: Akademisk Forlag A/S

URL: www.nys.dk



Søgbarhed

Artiklerne i de ældre NyS-numre (NyS 1-36) er skannet og OCR-behandlet. OCR står for 'optical character recognition' og kan ved tegngenkendelse konvertere et billede til tekst. Dermed kan man søge i teksten. Imidlertid kan der opstå fejl i tegngenkendelsen, og når man søger på fx navne, skal man være forberedt på at søgningen ikke er 100 % pålidelig.

Göteborgskorpusen för talspråk

(The Göteborg Spoken Language Corpus, GSLC)

JENS ALLWOOD, LEIF GRÖNQVIST, ELISABETH AHLSÉN OG MAGNUS GUNNARSSON

1. INLEDNING

Denna uppsats innehåller en beskrivning av talspråkskorpusen (GSLC) vid institutionen för lingvistik, Göteborgs universitet, samt en sammanfattning av de olika typer av analys och verktyg som har utvecklats för arbete med denna korpus. Arbete på korpusen inleddes under sent 1970-tal (det finns dock även material från 1960-talet) och har inkrementellt byggts på sedan dess. Idag innehåller korpusen ca. 1,3 millioner ord från omkring 25 olika sociala verksamheter. Korpusen har byggts upp för att tillgodose det växande intresset inom lingvistik för naturalistiska talspråksdata. En utgångspunkt är här att talspråk i stor utsträckning varierar i olika sociala verksamheter med avseende på uttal, ordförråd, grammatik och kommunikativa funktioner. Målsättningen för korpusen är att inkludera talspråk från så många typer av social verksamhet som möjligt för att få en mera fullständig förståelse av den roll språk och kommunikation spelar i mänskligt socialt liv. Denna typ av talspråkskorpus är fortfarande relativt unik, t.o.m. för engelska, eftersom många talspråkskorpora har insamlats för speciella syften såsom taligenkänning, fonetik, dialektal variation eller interaktion med ett datorstött dialogsystem. Oftast kommer också inspelningarna från en mycket begränsad verksamhet eller domän, se t.ex. Edinburgh Map Task (Isard och Carletta (1995), TRAINS (Heeman och Allen (1994), Waxholm, Blomberg m.fl. (1993) Jämfört med engelska korpora liknar Göteborgskorpusen kanske mest den nya zeeländska Wellington Corpus of Spoken New Zealand English (Holmes, Vine och Johnson 1998), men den har också gemensamma drag med BNC (British National Corpus) och London/Lund-korpusen (Svartvik 1990). Likheter finns också med den danska BySoc-korpusen

(Gregersen 1991, Henrichsen 1997). När det gäller inspelningar baseras korpusen till 50% på audio- och till 50% på videoinspelningar av naturalistiskt förekommande interaktion.

Inspelningarna har transkriberats enligt en transkriptionsstandard, GTS 6.2 (Nivre 1999b), (den har testats på kinesiska, arabiska, engelska, spanska, bulgariska och finska) och en språkspecifik del som gäller svenska – Modifierad Standard-Ortografi, MSO, f.n. version 6 (Nivre 1999a). Båda delarna har gått igenom 6 stora revisioner och flera mindre. För att förbättra reliabiliteten kontrolleras alla transkriptioner av en person utöver transkriptören. De kontrolleras också automatiskt så att deras format blir korrekt innan de inkluderas i korpusen. I MSO används standardortografi om det inte finns flera konventionella talspråksvarianter av ett ord. När det finns flera varianter hålls de isär grafiskt. Även om målet är att hålla transkriptionerna enkla, innehåller standarden talspråksdrag såsom kontrastiv betoning, överlapp och pauser. Den innehåller också procedurer för att anonymisera transkriptioner och för att introducera kommentarer gällande delar av transkriptionen.

Parallellt med att korpusen insamlats och transkriberats har kontinuerligt olika datorbaserad verktyg utvecklats för att underlätta arbetet med korpusen. Dessa beskrivs korfattat nedan. Genom att använda korpusen och dessa verktyg har vi kunnat göra olika typer av kvalitativ och kvantitativ analys, ett exempel på detta är en bok med jämförelser av frekvenser för svenskt tal- och skriftspråk (Allwood 1998). Boken innehåller ordfrekvenser både för ord skrivna i MSO-format och skrivna i standardortografiskt format. Den innehåller vidare statistik gällande ordklasser i tal och skrift, grundade på en automatisk probabilistisk taggning som ger 97% korrekt klassifikation. Korpusen har inte bara bearbetats automatiskt utan har också använts för olika typer av manuell kodning, t.ex. "kommunikationsreglering" (innefattande tvekljud, taländringar, återkoppling och turtagande), talakter, åtaganden, missförstånd etc. (Allwood 2001). Korpusen kan också utnyttjas för andra typer av kvalitativ analys, t.ex. för CA-relaterad sekventiell analys. Inspelningarna i korpusen digitaliseras kontinuerligt på digitala band eller CD med mpeg-kompression. Varje CD innehåller både transkriptioner och inspelningar.

2. GSLC OCH ANDRA KORPORA I GÖTEBORG

Talspråkskorpora vid institutionen för lingvistik vid Göteborgs universitet innehåller förutom GSLC flera andra typer av korpora, se tabell 1 nedan. Dessutom arbetar vi också med talspråkskorpora som insamlats av andra forskargrupper.

TABEL 1. Talspråkskorpora vid Göteborgs universitet, institutionen för lingvistik

- Göteborgskorpusen för talspråk – GSLC (kärnkorpusen -- vuxna förstaspråkstalare av svenska), 1,3 millioner ord
- Talare med afasi
- Barnspråkskorpus (svenska och andra nordiska språk), 0, 75 millioner ord inkluderande vuxna deltagare
- Utbildningsprocess, 416 longitudinella intervjuer, 2 millioner ord
- Talspråkskorpora med icke-svenska vuxna
  - Kinesiska (70 000 ord)
  - Bulgariska (25 000 ord)
  - Arabiska
  - Engelska (10 000 ord) + BNC
  - Finska
  - Italienska (3 000 ord)
  - Norska (140 000 ord)
  - Spanska
- Wizard-of-Oz och Bionisk korpus
- Interkulturell kommunikationskorpus

Det är kärnkorpusen (GSLC) vi kommer att fokusera på i denna artikel. I tabell 2 nedan presenterar vi några data om denna korpus. Som nämnts ovan är korpusen baserad på sociala verksamheter snarare än på t.ex. dialekter eller kategoriseringar av talare som socialklass eller kön. Emellertid kan omgrupperingar eller urval från korpusen göras på basis av sådana kriterier. De begränsningar som finns för våra möjligheter att skapa subkorpora är beroende av att vi inte alltid har den information som skulle behövas om individuella talare.

TABEL 2

Typ av social verksamhet	Antal inspel-	Genomt- snittlig ningar	Antal sektioner* antal talare	Ordföre- komster (in klusive pau- ser och kom- mentarer)	Hörbara ordföre- komster	Duration**
Auktion	2	6,0	111	26 776	26 459	3:14:11
Bussförare/ passagerare	1	33,0	20	1 360	1 345	0:13:33
Konsultation	16	3,0	239	34 865	34 285	2:44:25
Rättegång	6	5,0	79	33 401	33 261	3:58:33
Middag	5	8,0	30	30 738	30 001	2:49:54
Diskussion	34	5,8	255	240 426	237 583	17:19:24
Fabrik	5	7,4	48	29 024	28 860	2:19:47
Formellt möte	13	9,7	186	219 352	215 582	15:45:54
Hotell	9	19,2	183	18 950	18 137	6:47:50
Informellt samtal	22	4,4	152	94 490	93 436	7:48:41
Informationsservice	32	2,1	40	14 700	14 614	0:13:40
Intervju	58	2,9	1 031	396 758	393 907	30:34:27
Föreläsning	2	3,5	3	14 682	14 667	1:38:00
Marknad	4	24,2	38	12 581	12 175	2:18:37
Högmässa	2	3,5	10	10 273	10 234	1:10:45
Återbreättande av artikel	7	2,0	7	5 331	5 290	0:42:00
Rollspel	2	2,5	7	5 702	5 652	0:39:16
Affär	49	7,4	139	36 385	34 976	6:40:46
Uppgiftscen- trerad dialog	26	2,3	46	15 475	15 347	2:05:20
Terapi	2	7,0	8	13 841	13 529	2:04:07
Mässa	16	2,1	16	14 353	14 116	1:12:46
Resebyrå	40	2,7	112	40 370	40 129	5:53:57
Totalt	353	4,9	2 762	1 310 284	1 204 029	118:15:53

* En sektion är en längre fas av en verksamhet med ett distinktivt underordnat syfte. Bussförar-/passagerarinspelningarna har t.ex. 30 sektioner där varje sektion innehåller tal med en ny passagerare.

** För vissa inspelningar saknas uppgift om duration. Vi uppskattar att siffran ovan understiger den faktiska durationen med ungefär 30 timmar.

3. LAGRING

Omkring 50% av de 1,3 millioner ordförekomsterna är lagrade på audioband och resten finns på videoband (Umatic, VHS eller BetaCAM). För att kunna bevara inspelningarna, håller vi på att digitalisera dem genom att kopiera dem till digitala band. Ett mini-DV-band rymmer 60 minuter eller ett DVCam-band 180 minuter. Detta format kräver en snabb dator. Vid Mpeg-kompression har vi försökt att använda en konstant datahastighet på omkring 200 Kb per sekund. Detta ger en bra kvalitet och formatet kan användas på de flesta PC/Mac-maskiner.

4. BESKRIVNING AV KORPUSENS TRANSKRIPTIONSSTANDARD

Transkriptionsstandarden (GTS + MSO) vi har använt kan kanske lättast förklaras genom ett exempel.

EXEMPEL 1. Transkription enligt GTS + MSO

§1. Small talk
$D: säger du de{t} ä{r} de{t} ä{r} de{t} så besvärlit då
$P: ja ja
$D: m // ha / de{t} kan ju bli så se{r} du
$P: < jaha >
@ <ingressive>
$D: du ta{r} den på morronen
$P: nej inte på MORRONEN kan ja{g} ju tar allti en promenad på förmiddan [1 å0]1 då vill ja{g} inte ha [2 den]2 medicinen å0 sen nä ja{g} kommer hem möjligtvis
$D: [1 {j}a]1
$D: [2 nä]2

Exemplet visar följande egenskaper hos transkriptionsstandarden:

(i) Sektionsgränser markeras med paragraftecken (§) och delar upp en verksamhet i subaktiviteter. En läkar-patient-konsultation kan t.ex. ha följande subaktiviteter: (i) hälsning och introduktion, (ii) anledning till besöket, (iii) undersökning, (iv) diagnos, (v) förslag till behandling

(ii) Ord och mellanrum mellan orden

(iii) Dollartecken ($) följt av stor bodstav, följd av kolon (:) används för att indikera ny talare och ett nytt yttrande.

(iv) Dubbla snedstreck (//) används för att indikera pauser. Snedstreck /, // eller /// används för att indikera pauser av olika längd.

(v) Stora bokstäver används för att indikera kontrastiv betoning.

(vi) Ordindex används för att indikera vilket skriftspråksord som motsvarar den talspråksform som anges i transkriptionen. (å0) motsvarar skriftspråkets och. I de fall då talspråksvarianterna kan ses som förkortade former av skriftspråk, använder vi krullparenteser ({ }) för att visa vad den standardortografiska formen skulle vara, t.ex. de{t}.

(vii) Överlapp indikeras med hakparenteser ([ ]) med index, vilket tillåter disambiguering om flera talare överlappar samtidigt.

(viii) Kommentarer kan skrivas in genom att använda vinkelparenteser (< >) för att markera räckvidden på kommentaren i transkriptionen och (@< >) för att skriva in den aktuella kommentaren. Kommentarer kan t.ex. gälla händelser som är viktiga för interaktionen eller sådana fenomen som röstkvalitet och gester.

5. VERKTYG SOM HAR UTVECKLATS

Under den tid som korpusen har insamlats och transkriberats har många verktyg för att arbeta med korpusen utvecklats. Följande är fortfarande aktuella.

5.1. TRANSTOOL

TransTool (Nivre m.fl. 1998) är ett datorverktyg för att transkribera talspråk i enlighet med transkriptionsstandarden (Nivre 1999a,b). Det hjäl-

per den användare att transkribera korrekt och gör det lättare att hålla reda på index för överlapp och kommentarer (se Nivre et al 1998).

5.2 KORPUS-BROWSERN

Korpusbrowsern är ett verktyg som gör det möjligt att via internet söka på ord, ordkombinationer och fraser (som reguljära uttryck) i Göteborgskorpusen för talspråk. Resultaten kan presenteras som konkordanser eller listor av uttryck med så mycket kontext man vill ha och med direkta länkar till transkriptionen.

5.3 TRACTOR

TRACTOR är ett kodningsverktyg som gör det möjligt att skapa nya kodningsscheman och att koda transkriptioner. De segment i transkriptionen som kodas kan vara kontinuerliga eller diskontinuerliga och det är även möjligt att koda relationer. Ett kodningsschema kan representeras som ett träd med strängar på alla noder och löv och ett kodningsvärde är en "stig" genom trädet. Modellen liknar fil- och mappstrukturen på en datorhårddisk. Denna struktur gör det lättare att analysera kodningarna i ett prologsystem, men det är inte möjligt att ordna koderna eller att koda en kodning, eftersom en kod alltid består enbart av två diskontinuerliga intervall och ett kodat värde (Larsson 1997).

5.4 VISUALISERING AV KODER MED FRAMEMAKER

Vi har också skapat en verktygslåda som gör det möjligt att visualisera kodningsscheman och kodade värden med färg, fetstil, kursiv stil etc. direkt i transkriptionerna som ett FrameMaker-dokument. Olika delar av transkriptionen kan också markeras (eller uteslutas!) för att få en överskådlig bild utan de detaljer man kanske inte för tillfället är intresserad av (Grönqvist 1999).

5.5 TRASA

Om man har en korpus som är transkriberad enligt Göteborgsstandarden för transkription kan man genom att använda TraSA (Grönqvist 2000b)

relativt enkelt erhålla ett 30-tal statistiska mått för olika egenskaper, verksamheter, sektioner eller talare. Man kan t.ex. räkna antal ordförekomster, ordtyper, yttranden eller mer komplexa mått som ordrikedom.

5.6 SYNCTOOL

SyncTool (Nivre m.fl. 1998) är en prototyp för MultiTool nedan, som möjliggör synkronisering av transkriptioner med digitaliserade audio- och videoinspelningar. Den är också avsedd att vara ett "vyverktyg" som tillåter användaren att se transkriptionen och att spela upp det relaterade inspelade materialet, utan att behöva manuellt lokalisera de aktuella passagen i inspelningen.

5.7 ARBETE PÅ ETT SYNKRONISERINGSVERKTYG – MULTITOOL

Många av de ovan beskrivna verktygen skulle vara mer användbara om man kunde utnyttja de olika funktionerna simultant i ett verktyg. MultiTool är ett försök att bygga ett sådant verktyg för transkription och kodning av talspråk, liksom för "browsing", sökning och räkning. Systemet kan hantera ett godtyckligt antal talare, överlappande tal, hierarkiska kodningsscheman, diskontinuerliga kodningsintervall, relationer och synkronisering mellan kodningar och mediafiler (Grönqvist 2000a).

Den grundläggande idén är att samla all information i ett internt tillstånd som är en lågnivå-representation av alla typer av annotering (kodning), inklusive transkription. Tillståndet innehåller de abstrakta objekten kodning och synkroniseringar. Detta är de typer av grundläggande information datorprogrammet behöver. För användare som utnyttjar audio- och videoinspelningarna i korpusen är transkriptionerna enbart en kodning av inspelningarna. En viktig detalj är att alla vyer (t.ex "partitur" eller andra vyer av transkriptionen, vyer av kodningar och akustisk analys, liksom även videofiler) som är kopplade till samma tidpunkt kan synkroniseras för att visa samma sekvens från olika perspektiv närhelst en användare utnyttjar en av dem. Det interna tillståndet innehåller all information, så det är möjligt att ha flera olika vyer på samma sekvens i en dialog. Förändringar av något i en vy kommer omedelbart att förändra det inre tillståndet och som en konsekvens härav de andra vyerna.

MultiTool är skrivet i JAVA + JMF, vilket gör programmet förhållandevis plattformsoberoende och eftersom interpretatorerna snabbt blir mer effektiva, kommer troligen prestanda att bli tillräckligt bra på alla viktiga plattformar inom den närmaste framtiden. En ny egenskap vi håller på att lägga till är import- och exportformat för våra lokala transkriptionsformat, TRACTOR-filer och troligen också för CA- transkriptioner (CA = Conversation Analysis).

Vår ambition är att de nya versionerna av MultiTool i framtiden för många användare kommer att ersätta de olika verktygen vi har beskrivit ovan. Emellertid kommer TraSA och Korpusbrowsern fortfarande att behövas när man arbetar på stora delar av korpusen samtidigt. Med adekvata import/exportfunktioner kommer olika användare att kunna använda sina egna transkriptions- och kodningsformat i MultiTool. På så sätt hoppas vi att MultiTool kommer att utgöra en god basnivå för analys av mutlimodala talspråkskorpora: transkription, annotering/kodning, konversion, söklning, räkning, "browsing" och visualisering. För användare med andra intressen finns dock bättre verktyg, som t.ex. Waves för fonetiker och MediaTagger för enklare kodningar av audio/videofiler.

6. TYPER AV KVANTITATIV ANALYS

På grundval av den information som ges av transkriptioner enligt Göteborgsstandarden har vi definierat en uppsättning egenskaper som kan härledas automatiskt ur transkriptionerna. Några av dessa egenskaper är följande (se Allwood och Hagman 1994, Allwood 1996):

(i) **Volym:** Volym omfattar mått som antal ord, ordlängd, pauser, betoning, yttranden och turer relativt talare, verksamhet och subaktivitet.

(ii) **Kvoter:** Ifrån volymmåtten kan sedan olika kvoter räknas fram. T.ex.:
MLU = ord / yttrande
% pauser = 100 x pauser / (ord + pauser)
% betoning = 100 x betonade ord / ord
% överlapp = 100 x överlappade ord / ord
hastighet = ord / duration

Alternativt kan pauser, betoning och överlapp beräknas per yttrande. Alla dessa kvoter kan sedan relateras till talare, verksamhet eller subaktivitet (sektion).

(iii) **Speciella deskriptorer:** Ett exempel på en "speciell deskriptor" är "ordrikedom", som kan mätas genom ordförekomst / ordtyp. Guiraud, über, Herdan eller "teoretisk vokabulär", cf. Van Hout och Rietveld (1993). Andra deskriptorer som vi har konstruerat är "stereotypiskhet, som räknar ut hur ofta ord och fraser upprepas i en verksamhet, "verbal dominans" och verbal jämlikhet", "livlighet" och "försiktighet" samt "överlapp" i olika yttrandepositioner.

(iv) **Lemma:** Vi har också implementerat en enkel "stam"-algoritm som gör det möjligt för oss att gruppera regelbundet böjda former med sin ordstam.

(v) **Ordklasser:** Orden i korpusen kan tilldelas ordklasser genom att använda en sannolikhetsbaserad statistisk (Viterbi-trigram) ordklasstaggare som har anpassats till talspråk. Genom att använda denna har ordklasstaggning gjorts för hela GSLC (ungefär 1,3 millioner transkriberade ord). Korrektheten är ungefär 97% (cf. Nivre och Grönqvist 2001). Ord som taggats för ordklass kan sedan tilldelas talare, verksamhet och subaktivitet.

(vi) **Kollokationer:** Alla talare, verksamheter och subaktiviteter kan beskrivas med avseende på vilka kollokationer som förekommer. Dessa kan sorteras efter frekvens, efter förekomst som fullständiga yttranden eller efter "mutual information" (Manning och Schütze 1999).

(vii) **Frekvenslistor:** Frekvenslistor kan göra för ord, lemman, ordklasser, kollokationer och yttrandetyper.

(viii) **Sekvenser av ordklasser:** Yttranden av olika längd kan beskrivas med avseende på vilka ordklassekvenser de innehåller. Detta tillåter en första analys av grammatiska skillnader mellan talare, verksamheter och subaktiviteter.

(ix) **Likheter:** Likheter mellan verksamheter kan fångas genom att analysera i hur stor utsträckning ord och kollokationer delas mellan verksamheter.

Validitets- och reliabilitetskontroll görs manuellt av alla automatiska mått.

7. TYPER AV KVALITATIV ANALYS

7.1 ÖVERSIKT

För att öka reliabiliteten i kodning, har kvalitativ analys i Göteborg ofta resulterat i utvecklandet av kodningsscheman, dvs. scheman för annotation ovanpå transkriptioner. De kodningsscheman som utvecklats i Göteborg kan jämföras med andra scheman och då kan vi se att några av dessa ligger ovanpå transkription, t.ex. DAMSL (Core and Allen 1997) and DRI, medan andra är integrerade med transkriptionsstandarden, t.ex. uppmärkningsramen i MATE (Dybkjaer m.fl. 1998). En rättvis jämförelse mellan de viktigaste, för att inte säga alla scheman ligger utanför ramarna för denna redogörelse. De kodningsscheman som presenteras nedan reflekterar således de intresseområden Göteborgs-gruppen har fokuserat på. Den underliggande transkriptionsstandarden begränsar på ett naturligt sätt finkornigheten för alla nya kodningsscheman, men de två kodningsverktyg som utvecklats i Göteborg, MultiTool och TRACTOR, är avsedda att vara så oberoende av alla individuella kodningsscheman och transkriptionsstandarder som möjligt. Följande lista ger en översikt av kodningsscheman från Göteborg (cf. Allwood 2001).

Kodning relaterad till:

1. Social verksamhet och kommunikativa akter
   1.1 Social verksamhet
   1.2 Kommunikativa akter
   1.3 Expressiva och evokativa funktioner
   1.4 Förpliktelser (åtaganden)

2. Kodning relaterad till kommunikationsreglering
   2.1 Återkoppling (feedback)
   2.2 Tur- och sekvensreglering
   2.3 Egen kommunikationsreglering

3. Grammatisk kodning
   3.1 Ordklasser (automatisk, probabilistisk)
   3.2 Maximala grammatiska enheter

4. Semantisk kodning

Kontroll av reliabilitet är planerad att inkluderas i utvecklingen av alla kodningsscheman. Hittills har sådan kontroll gjorts av kodning för "återkoppling" och "egen kommunikationsreglering" (med hjälp av Cohens kappa).

7.2 BIDRAG, YTTRANDEN OCH TURER

I enlighet med Grice (1975), Allwood, Nivre och Ahlsén (1990) och Allwood (2000), antas de grundläggande enheterna i dialog vara gestuella eller vokala bidrag från deltagarna. Termen bidrag används istället för yttrande, när vi vill inkludera inte bara muntlig vokal input till kommunikationen utan också gester eller skriftlig input. Verbala bidrag kan bestå av enstaka morfem eller vara flera satser långa. Termen tur används för "rätten att bidra" snarare än för det bidrag som produceras genom användande av denna rätt. Man kan "göra ett bidrag" utan att "ha turen" och man kan "ha turen" utan att använda den för ett aktivt bidrag. Ett exempel på detta ges nedan, där B:s första bidrag innebär givande av positiv återkoppling utan att ha turen (hakparenteser indikerar överlapp) och B:s andra bidrag innebär att han/hon under sin tur är tyst och inte gestikulerar.

A: titta glass [vill] du ha en glass
B1: [ja]
B2: (tystnad och ingen handling)

Bidrag, yttranden och turer kodas inte eftersom de kan fås direkt ur GTS, den Göteborgska transkriptionsstandarden.

7.3 KODNING RELATERAD TILL SOCIAL VERKSAMHET OCH KOMMUNIKATIVA AKTER

7.3.1 Social verksamhet

Varje transkription är länkad till en databaspost och ett "huvud" (header) som innehåller information om:

(i) Syfte(n), funktion(er) och procedurer i verksamheten

(ii) Verksamhetens roller

(iii) Artefakterna, dvs. objekt, möbler, instrument och media som utnyttjas i verksamheten

(iv) Den sociala och fysiska omgivningen

(v) Data om deltagarna (anonymiserade), såsom ålder, kön, dialekt och etnicitet

Dessutom anges de viktigaste subaktiviteterna för varje verksamhet.

7.3.2 Kommunikativa akter

Varje bidrag kan kodas med hänsyn till vilka kommunikativa akter den innehåller simultant eller sekventiellt. De kommunikativa akterna finns på en lista som kan utvidgas. De flesta typer har idag definitioner och operationalisering. Några av de typer som används ofta är följande: Uppmaning, Påstående, Tvekan, Fråga, Svar, Specifikation, Konfirmation (Bekräftelse), Affirmation (Bekräftelse), Avslutande av interaktion, Avbrott, Slutsats och Erbjudande.

7.3.3 Expressiva och evokativa funktioner

I enlighet med Allwood (1976, 1978, 2000) anses varje bidrag ha en expressiv och en evokativ funktion. Dessa funktioner explicitgör några av de funktioner som impliceras av kodningen av kommunikativa akter. Den expressiva funktionen lter sändaren uttrycka trosuppfattningar och

andra kognitiva attityder och känslor. Vad som "uttrycks" består av en kombination av reaktioner på föregående bidrag och nya initiativ. Den evokativa funktionen är den reaktion sändaren avser att "framkalla" hos lyssnaren. På så sätt är den evokativa funktionen hos ett påstående normalt att "framkalla" samma uppfattning som "uttryckts" i påståendet hos lyssnaren. Den evokativa funktionen hos en fråga är att framkalla ett svar, medan den evokativa funktionen hos en uppmaning är att framkalla en önskad handling.

7.3.4 Förpliktelser (åtaganden)

Om dialog och kommunikation skall fungera på ett kooperativt sätt, oavsett om detta sker som medel för en annan verksamhet eller ej, nödvändiggörs vissa förpliktelser och åtaganden för både talare (sändare) och lyssnare (mottagare). Med avseende på både expressiva och evokativa funktioner, bör sändaren ta hänsyn till mottagarens perceptuella, kognitiva och beteendemässiga förmåga och bör inte vilseleda, skada eller onödigtvis inskränka mottagarens frihet. Mottagaren bör tillmötesgå med en värdering av huruvida hon/han kan höra, förstå och utföra det som ges av sändarens evokativa avsikter och signalera detta till sändaren. Sändarens och mottagarens förpliktelser och åtaganden kan summeras på följande sätt (se också Allwood 1994):

Sändaren (åtaganden): 1. Uppriktighet, 2. Motivation, 3. Hänsyn (se Allwood 1976).

Mottagaren (förpliktelser): 1. Värdering, 2. Rapport, 3. Handling.

7.4 KODNING RELATERAD TILL KOMMUNIKATIONSREGLERING

7.4.1 Inledning

Termen "kommunikationsreglering" syftar på de medel som talare kan använda för att reglera interaktionen eller sin egen kommunikation. Det finns tre kodningsscheman som är relaterade till kommunikationsreglering (se Allwood m.fl. 1999):

1) Kodning av återkoppling
2) Kodning av tur- och sekvensreglering
3) Kodning av egen kommunikationsreglering

7.4.2 Kodningsschema för återkoppling

En återkopplingsenhet kan beskrivas som "en maximal kontinuerlig utsträckning av ett yttrande (förekommande självständigt eller som del av ett längre yttrande), vars primära funktion är att ge och/eller framkalla återkoppling rörande kontakt, perception, förståelse och acceptans av evokativ funktion" (Allwood 1993). Alla återkopplingsenheter kodas med avseende på "Struktur", "Position/Status", och "Funktion". Att koda struktur betyder att koda grammatisk kategori (satsdel, fras eller mening) och även "strukturella operationer". "Strukturella operationer" indelas i "fonologiska", "morfologiska" och kontextuella" operationer, vilka var och en har olika värden.

7.4.3 Kodning av tur- och sekvensreglering

Tur- och sekvensreglering omfattar följande fenomen:

(A) Överlapp och avbrott: Överlapp kodas i transkriptionerna och kan extraheras automatiskt. Avbrott är en kod för de överlapp som syftar till att eller lyckas byta ämne eller ta turen från en annan talare.

(B) Avsedd mottagare: Denna typ av kodning har 4 självförklarande värden:
(i) en viss deltagare

(ii) en viss grupp av deltagare

(iii) alla deltagare

(iv) ingen annan deltagare (att tala till sig själv)

(C) Markerande av inledande och avslutande av subaktiviteter och/eller interaktionen som helhet.

7.4.4 Kodningsschema för egen kommunikations-reglering (EKR)

EKR betyder "Egen kommunikations-reglering" och står för processer som talare använder för att reglera sina egna bidrag i kommunikativ interaktion. Att koda EKR-funktion innebär att klassificera om EKR-enheten är:

- Val-relaterad – hjälper talaren att vinna tid för processer som berör fortlöpande val av innehåll och typer av strukturella uttryck, eller:
- Ändrings-relaterad – hjälper n att ändra innehåll, struktur eller uttryck som redan producerats.

EKR-enheter kodas också med avseende på det EKR-relaterade uttryckets struktur. Denna struktur kan indelas i "grundläggande EKR-drag", "grundläggande EKR-operationer" och "komplexa EKR-operationer". Pauser, enkla EKR-uttryck som tvekljud etc. och explicita EKR-fraser räknas som grundlläggande EKR-drag. Grundläggande EKR-operationer är: "förlängning av kontinuanter", "själv-avbrott" och "självupprepning". Kategorin "komplexa EKR-operationer" står för olika sätt att modifiera den språkliga strukturen. EKR-kodningsschemat beskrivs i Allwood m.fl. (1997).

7.5 GRAMMATISK KODNING

Det finns också möjligheter att koda grammatisk struktur. En av dessa är den ovannämnda automatiska ordklasstaggningen. En annan är kodning av "maximala grammatiska enheter" – ett kodningsschema som finns beskrivet i Allwood (2001). När man kodat "maximala grammatiska enheterd" bör man i första hand försöka hitta så stora enheter som möjligt, den största enheten är härvidlag "fullständiga satser". Satser kan subklassificeras genom att använda schemat "satser". I talspråk finns det många yttranden som inte är satser, så i andra hand bör man försöka hitta "fullständiga fraser". Dessa bör kodas med schemat "fraser". Om det inte är möjligt att finna vare sig fullständiga satser eller fullständiga fraser, kodas

enskilda ord med schemat "ordklasser". Vart och ett av de tre nämnda schemana innehåller flera underkategorier.

8. SLUTSATSER OCH FRAMTIDA ARBETE

I denna uppsats har vi beskrivit en del av det arbete som gjorts vid institutionen för lingvistik vid Göteborgs universitet för att samla, transkribera och lagra talspråksmaterial. Vi har också beskrivit några av de verktyg som har utvecklats för att underlägga arbetet med att analysera data, både automatiskt och manuellt. Slutligen har vi beskrivit några av de resultat vi hittills erhållit. Framtida arbete kommer att inkludera en inkrementell utvidgning av korpusen både för att få data från nya social verksamheter och för att utjämna storleken på inspelat och transkriberat material från olika verksamhetstyper. Vi kommer också att göra flera ansträngningar att göra korpusen mera multimodal genom att göra de audio- och videoinspelningar som transkriptionerna bygger på mera tillgängliga. Arbete på verktyg för att analysera korpusen kommer att fortsätta. Det mest omedelbara målet är att komplettera MultiTool, vilket förhoppningsvis kommer at ge oss bättre möjligheter att arbeta med multimodala data. Parallellt med detta kommer arbete på kvalitativ och kvantitativ analys att fortsätta. Ett ambitiöst mål är att arbeta mot en grammatisk beskrivning av talspråk och mot en systematisk beskrivning (även om detta kanske inte skall vara en grammatik) av multimodal ansikte-mot-ansikte-kommunikation.

Jens Allwood
Institutionen för Lingvistik,
Göteborg Universitet
email: jens@ling.gu.se

Leif Grönqvist
Institutionen för Lingvistik,
Göteborg Universitet
email: leifg@ling.gu.se

Elisabeth Ahlsén
Institutionen för Lingvistik,
Göteborg Universitet
email: elisa@ling.gu.se

Magnus Gunnarsson
Institutionen för Lingvistik,
Göteborg Universitet
email: mgunnar@ling.gu.se

LITTERATUR

Allwood, J. (1976): *Linguistic Communication as Action and Cooperation.* Gothenburg Monographs in Linguistics 2. Göteborgs universitet, institutionen för lingvistik.

Allwood, J. (1978): On the Analysis of Communicative Action. M. Brenner (red.): *The Structure of Action*: 168-191. Oxford: Basil Blackwell.

Allwood, J. (1993): Feedback in Second Language Acquisition. C. Perdue (red.): *Adult Language Acquisition. Cross Linguistic Perspectives*, Vol. II: 37-51. Cambridge: Cambridge University Press.

Allwood, J. (1994): Obligations and Options in Dialogue. *Think*, Vol 3, May: 9-18. ITK, Tilburg University.

Allwood, J. (red.) (1996 and later editions): *Talspråksfrekvenser*, Ny och utvidgad upplaga. Gothenburg Papers in Theoretical Linguistics S21. Göteborgs universitet, institutionen för lingvistik.

Allwood, J. (1998): Some Frequency based Differences between Spoken and Written Swedish. T. Haukioja (red.): *Proceedings of the 16th Scandinavian Conference of Linguistics*: 18-29. Turku University, Department of Linguistics.

Allwood, J. (2000): An Activity Based Approach to Pragmatics. H. Bunt, & B. Black (red.): *Abduction, Belief and Context in Dialogue; Studies in Computational Pragmatics*: 47-80. Amsterdam: John Benjamins.

Allwood, J. (red.) (2001): *Dialog Coding – Function and Grammar: Göteborg Coding Schemas.* Gothenburg Papers in Theoretical Linguistics; GPTL 85. Göteborgs universitet, institutionen för lingvistik.

Allwood, J. & Hagman, J. (1994): Some Simple Measures of Spoken Interaction. F. Gregersen & J. Allwood (red.): *Spoken Language, Proceedings of the XIV Conference of Scandinavian Linguistics*: 3-22.

Allwood, J., Ahlsén, E., Nivre, J. & Larsson, S. (2001): Own communication management. Allwood, J. (red.) (2001): *Dialog Coding – Function and Grammar: Göteborg Coding Schemas*: 45-52. Gothenburg Papers in Theoretical Linguistics; GPTL 85. Göteborgs universitet, institutionen för lingvistik.

Allwood, J., Nivre, J. & Ahlsén, E. (1990): Speech Management: On the Non-Written Life of Speech. *Nordic Journal of Linguistics* 13: 3-48.

Blomberg, M., Carlson, R., Elenius, K., Granström, B., Gustafson, J, Hunnicutt, S., Lindell, R. & Neovius, L (1993): An experimental dialogue system: WAXHOLM. *Proceedings of EUROSPEECH 93*: 1867-1870.

Core, M. G. & Allen, J. F. (1997): Coding Dialogs with the DAMSL Annotation Scheme. *Working Notes of AAAI Fall Symposium on Communicative Action in Humans and Machines.* Boston, MA, November 1997.

Dybkjær, L., Bernsen, N.O., Dybkjær, H., McKelvie, D. & Mengel, A. (1998): *The MATE Markup Framework.* MATE Deliverable D1.2, November 1998.

Gregersen, F. (1991): *The Copenhagen Study in Urban Sociolinguistics* I-II. København: Reitzel.

Grice, H.P. (1975): Logic and conversation. *Syntax and Semantics,* Vol. 3: P. Cole & J.L. Morgan (red.): *Speech Acts*: 41-58. New York: Seminar Press.

Grönqvist, L. (1999): *Kodningsvisualisering med Framemaker.* Göteborgs universitet, institutionen för lingvistik.

Grönqvist, L (2000a): *The MultiTool User's Manual. A tool for browsing and synchronizing transcribed dialogues and corresponding video recordings.* Göteborgs universitet, institutionen för lingvistik.

Grönqvist, L. (2000b): *The TraSA v0.8 Users Manual. A user friendly graphical tool for automatic transcription statistics.* Göteborgs universitet, institutionen för lingvistik.

Heeman, P.A. & Allen, J.F. (1994): The TRAINS 93 Dialogues. *TRAINS Technical Note* 94-2.

Henrichsen, P.J. (1997): *Talesprog med Ansigtsløftning.* IAAS, Univ. of Copenhagen, Instrumentalis 10/97.

Holmes, J., Vine, B. & Johnson, G. (1998): *Guide to the Wellington Corpus of Spoken New Zealand English.* Victoria University of Wellington, Wellington.

Hout, R. v. & Rietveld, T. (1993): *Statistical Techniques for the Study of Language and Language Behaviour.* Berlin & New York: Mouton de Gruyter.

Isard, A. & Carletta, J. (1995): *Transaction and action coding in the Map Task Corpus.* Research Paper HCRC/RP-65.

Larsson, S. (1997): *TRACTOR v1.0b1 användarmanual.* Göteborgs universitet, institutionen för lingvistik.

Manning, C. D. & Schütze, H. (1999): *Foundations of Statistical Natural Language Processing.* Boston, Mass.: The MIT Press.

Nivre, J. (1999a): *Transcription Standard. Version 6.2.* Göteborgs universitet, institutionen för lingvistik.

Nivre, J. (1999b): *Modifierad StandardOrtografi (MSO) Version 6.* Göteborgs universitet, institutionen för lingvistik.

Nivre, J., Tullgren, K., Allwood, J., Ahlsén, E., Holm, J., Grönqvist, L., Lopez-Kästen, D. & Sofkova, S. (1998): *Towards multimodal spoken language corpora: TransTool and SyncTool.* Proceedings of ACL-COLING 1998, June 1998.

Nivre, J. & Grönqvist, L. (2001): Tagging a corpus of Spoken Swedish. *International Journal of Corpus Linguistics.*

Svartvik, J. (red.) (1990): *The London Corpus of Spoken English: Description and Research.* Lund Studies in English 82. Lund University Press.