

Det talade språket i ett longitudinellt perspektiv

KJELL HÄRNQVIST

Institutionen för pedagogik och didaktik, Göteborgs universitet

Sammanfattning: *Inom projektet Långtidseffekter av utbildning (LING) gjordes under 1982 en serie personliga intervjuer med cirka 500 personer födda 1948 som 1961 hade deltagit i det så kallade individualstatistik-projektets insamling av basdata i Årskurs 6. Denna artikel sammanfattar resultaten från en analys av ordförrådet hos de intervjuade relaterat till ett antal bakgrundsuppgifter från 13 års ålder och framåt. En detaljerad redovisning publiceras i tidskriften Computers and the Humanities (maj 2003). Den studerar skillnader mellan män och kvinnor, mellan lägre och högre socialgrupper och mellan lägre och högre utbildningsgrupper. De mest intressanta fynden avser sambandet mellan språklig nivå vid 13 och 32 års ålder och användningen av olika ordklasser i intervjun vid 34 års ålder.*

Det är nu otroliga 55 år sedan min första uppsats kom i tryck. Den hette »Yrkesplaner hos stockholmsgymnasister» och rapporterade en kortlivad longitudinell undersökning. Och på den vägen är det vanligtvis fortfarande när jag publicerar något, fast projekten efterhand har blivit mer omfattande och långlivade. Jag säger ibland att jag gör om samma undersökning minst en gång per decennium, bara med lite mer raffinerade metoder men tyvärr inte med så annorlunda resultat. Rubriken på mitt bidrag, »Det talade språket i ett longitudinellt perspektiv», kan tyckas innebära ett trendbrott, men longitudinell i rubriken antyder att skillnaden mot tidigare kanske inte är så stor. Jag har helt enkelt bytt ut de vanliga måtten på pedagogiska långtidseffekter mot andra mått hämtade från datalingvistikens revir, för övrigt en mycket spännande bekantskap.

FÖRST NÅGOT OM BAKGRUNDEN

Under de första åren av 1980-talet bedrev jag tillsammans med Ulf Christianson det så kallade LING-projektet (Långtidseffekter av utbildning). Detta projekt var en uppföljning av personer födda 1948. Uppföljningen omfattade bland annat cirka 500 intervjuer om studie- och yrkesval för att fördjupa det vi hade belyst med en stor enkät till 8 500 deltagare. Intervjuerna registrerades ordagrant för databehandling och en mindre rapport om ordförrådet i

intervjuerna gjord tillsammans med Ulf Christianson och Jan-Gunnar Tingsell (Härnqvist, Christianson & Tingsell 1985) kom ut 1985. Sedan har det språkliga materialet legat tämligen outnyttjat tills jag för ett par år sedan beslöt att ta upp intervjusvaren till ny behandling, därtill stimulerad av nya grepp på studiet av texter. Jag fick då också hjälp av Daniel Ridings som bidrog med ett program för automatisk klassificering av ordklasser. Nu har vi färdigställt en längre uppsats (Härnqvist, Christianson, Ridings & Tingsell 2003) som förhoppningsvis skall platsa i en av de internationella tidskrifterna inom språklig databehandling. Det är några resultat ur denna uppsats som jag skall redovisa, men först något om de data vi disponerar.

Lingvisterna använder termen corpus för att beteckna en databas byggd på löpande text eller transkriberade samtal. Befintliga skriftspråskorpusar är numera enormt stora. De bygger till stor del på tidningstext som tas direkt ur tryckunderlaget. Däremot är talspråskorpusar av större format sällsynta. Våra intervjupersoner producerade tillsammans cirka 1,8 miljoner ord, vilket placerar denna korpus bland de största av sitt slag internationellt sett. Därtill är den unik i flera hänseenden: den bygger på en representativ grupp av individer och de individer som intervjuas vid 34 års ålder kan följas bakåt i våra register till 13 års ålder. Det är sådant som gör att vi bedömer att våra resultat kan vara intressanta för en internationell publik.

I analyserna hittills har vi använt två olika uppsättningar av variabler som baseras på ordförrådet i intervjuerna. Den ena bygger på antalet ord, antalet olika ord, antalet ord med olika frekvens, längd och dylikt samt uppdelningen i stycken och förekomsten av avbrott. Den andra uppsättningen av variabler avser ordens fördelning på ordklasser. Som bakgrundsvariabler använder vi kön, socialgrupp, utbildningsnivå samt mått på språklig förmåga dels vid 13 års ålder, dels enligt självbedömningar i enkäten vid 32 års ålder. Jag skall nu ge några exempel på sådana jämförelser mellan bakgrund och ordförrådsvariabler.

SKILLNADER MELLAN MÄN OCH KVINNOR

Tabell 1. Skillnader avseende ordförrådet mellan män och kvinnor.

<i>Antal</i>	<i>Män-Kvinnor/</i>		
	<i>Män</i>	<i>Kvinnor</i>	<i>Spridning</i>
Ord	4615	4083	0.31
Olika ord	846	753	0.40
Ord med f=1	494	432	0.43
Ord med >10 bokstäver	79	63	0.43
Avbrott	312	262	0.35

Män ligger högre inte bara i de fem variabler som finns i tabellen ovan utan i alla frekvensvariabler och skillnaden är i allmänhet av storleken 30–40% av spridningen, vilket är betydande. Det är också skäl att notera att skillnaderna

är större i de mer kvalitetsbetonade variablerna *olika ord*, *ord som bara förekommer en gång* och *långa ord*, större alltså än i det sammanlagda antalet ord och i avbrott. Detta stämmer med internationella resultat som säger att män talar mer än kvinnor i formella situationer, medan populäruppfattningen att kvinnor talar mer än män gäller bara i informella sammanhang.

Våra intervjuer kan uppfattas som förhållandevis formella. Tyvärr kan vi inte kontrollera om det finns ett samspel mellan intervjuarens och respondentens kön ty bara tre av våra 16 intervjuare var män. Samma skillnader mellan könen slår igenom för det absoluta antalet ord i en viss ordklass. Men relativt sett använder kvinnorna oftare verb och pronomen än män, och dessa ordklasser har som vi senare skall se en lägre status än till exempel substantiv och prepositioner. På det hela taget är skillnaderna i ordanvändning större mellan vuxna män och kvinnor än skillnader i verbal förmåga såsom de visar sig i begåvnings- och kunskapstest på skolstadiet. Där är flickorna vanligtvis överlägsna. Troligen är det ett förhållande som ändrar sig över åren under påverkan av till exempel yrkesskillnader mellan män och kvinnor.

SKILLNADER MELLAN SOCIALA GRUPPER

Tabell 2. Skillnader avseende ordförrådet mellan sociala grupper.

Antal	Socialgrupp			Medel-Låg/ Spridning	Hög-Medel/ Spridning
	Låg	Medel	Hög		
Ord	4135	4272	5113	0.08	0.50
Olika ord	763	791	921	0.12	0.56
Ord med f=1	440	458	540	0.13	0.57
Ord med >10 bokstäver	64	71	91	0.17	0.54
Avbrott	264	282	360	0.12	0.54

Här delas materialet upp i tre klasser enligt föräldrarnas socialgrupp: hög, mellan och låg ungefär som i den gamla skalan av socialgrupperna ett till tre. Grupperna skiljer sig tydligt åt, men skillnaden går huvudsakligen mellan hög- och mellangrupperna där den är större än 50% av spridningen. Detta gäller som sagt föräldragenerationen där skillnaden mellan hög socialgrupp och övriga grupper sammanhänger med om någon av föräldrarna har en utbildning utöver den obligatoriska. En högutbildad familj kan mycket väl tänkas påverka språkutvecklingen under uppväxtåren och ge varaktiga effekter på den vuxnes ordförråd och sätt att hantera språket, antagligen både i tal och skrift.

SKILLNADER MELLAN UTBILDNINGSGRUPPER

Här är det fråga om respondenternas egen utbildning efter grundskolan upp till 30-årsåldern. Där går den stora klyftan mellan dem som bara har gått i

folkskola eller enhetsskola och dem som har gymnasial utbildning eller mera. Kom ihåg att de flesta 48:or gick i skolan före de stora skolreformerna då övergången till realskola var den viktigaste brytpunkten. När det gäller den relativa användningen av ordklasser ligger låggruppen högst i interjektioner (utropsord) och pronomen som tycks vara ordklasser med låg status.

Tabell 3. Skillnader avseende ordförrådet mellan utbildningsgrupper.

<i>Antal</i>	<i>Utbildningsnivå</i>			<i>Medel-Låg/ Spridning</i>	<i>Hög-Medel/ Spridning</i>
	<i>Låg</i>	<i>Medel</i>	<i>Hög</i>		
Ord	3839	4352	4805	0.30	0.27
Olika ord	697	817	881	0.52	0.28
Ord med f=1	399	475	514	0.53	0.27
Ord med >10 bokstäver	49	74	89	0.67	0.41
Avbrott	251	296	313	0.31	0.12

Användningen av långa ord (ord med mer än 10 bokstäver enligt vår gränsdragning) är den variabel som skiljer mest mellan utbildningsnivåerna.

Sammanfattningsvis för de tre gruppindelningarna tycks gälla att det talade språket såväl vad gäller kvantitet som kvalitet avspeglar dominansskillnader i samhället; mellan könen, mellan socialgrupperna och mellan utbildningsnivåerna. De intervjuade i de dominanta grupperna använder mer ord men framför allt mer olika ord och mer långa ord. De talar mer sammanhängande, gör avbrott när det passar dem själva och använder ordklasser som ger högre precision åt deras svar.

SKILLNADER MELLAN INDIVIDER

Låt oss därefter titta på hur användningen av olika ordklasser förhåller sig till individernas språkliga förmåga, dels vid 13 års ålder (Sp13), dels vid 32 års ålder (Sp32). För 13-åringarna är det medeltalet av ett motsatsprov ur vårt intelligenstest, standardproven i läsning, skrivning och engelska samt betygen i svenska och engelska i Årskurs 6. För 32-åringarna är det medeltalet av deras egna skattningar i enkäten av förmågan att läsa, skriva, lyssna och tala på svenska och engelska.

Tabell 4. Samband mellan ordklasser och språklig förmåga vid 13 och 32 års ålder.

<i>Positivt</i>	<i>Språk 13</i>	<i>Språk 32</i>	<i>Negativt</i>	<i>Språk 13</i>	<i>Språk 32</i>
Substantiv	0.31	0.41	Adverb	-0.19	-0.34
Adjektiv	0.28	0.38	Pronomen	-0.23	-0.31
Konjunktion	0.32	0.30	Interjektion	-0.30	-0.29
Determinativ	0.19	0.26	Verb	-0.18	-0.18
Preposition	0.16	0.20	Partikel	-0.12	-0.04

Tabellen visar korrelationerna mellan språkfärdigheter och användningen av olika ordklasser. Ordklasserna delar upp sig på ett mycket tydligt och konsistent sätt i en grupp med positiva korrelationer och en grupp med negativa korrelationer. Uppdelningen är densamma för både 13- och 32-årsmått. Särskilt intressant är kontrasten mellan substantiv och pronomen för att referera till personer, saker och händelser. Substantiven dominerar hos dem med högre språklig förmåga och pronomenen hos de övriga. En motsvarande kontrast finns belagd i engelska och amerikanska undersökningar där man funnit en faktor som differentierar mellan inriktning på information och inriktning på interaktion. Skriftspråk ligger vanligen högt på den informativa sidan och talspråk på interaktionssidan. I denna undersökning kan man konstatera att intervjuspråket har en kontrast inom sig som liknar skillnaden mellan skrift och tal och som korrelerar med individernas språkliga färdighet. Somliga talar som en bok.

JÄMFÖRELSER MED DENNA OCH ANDRA CORPUSAR

Tabell 5. Jämförelser med andra corpusar.

<i>Andelen i ordklasser</i>	<i>Corpus</i>		
	<i>Härnqvist m fl</i>	<i>Allwood</i>	
	<i>Tal</i>	<i>Tal</i>	<i>Skrift</i>
med positiva korrelationer	36%	35%	47%
med negativa korrelationer	64%	65%	51%

De ordklasser som korrelerar positivt med språklig förmåga omfattar tillsammans 36% av alla ord och de negativa återstående 64%. Denna relation överensstämmer nästan exakt med vad Jens Allwood (1999) funnit i sin talspråkscorpus. I hans skriftspråkscorpus som bygger på romaner och tidningstext är relationen däremot så nära 50/50 man kan komma. Själv hade jag skrivit en teoretisk text om samma ämne som intervjuerna, nämligen studie- och yrkesval. Där blev relationen 70/30, alltså en extrem skriftspråklig typ av fördelning.

Åter till korrelationerna. Andelen i »positiva» ordklasser korrelerade 0.38 med språkförmågan vid 13 års ålder och 0.47 med självbedömningarna av språkförmåga vid 32 års ålder. De »negativa» ordklasserna hade lika höga korrelationer men med ett minustecken framför. Det som kan synas anmärkningsvärt här är att sambandet över 21 års tid är nästan lika högt som sambandet över två år. Hur kan det komma sig?

Den frågan kan man belysa med en så kallad stig- eller pathanalys där man lägger de olika påverkande faktorerna i en kronologisk ordning och ritar upp ett flödesdiagram som skiljer isär direkta och indirekta effekter. Metodiken bygger på multipel stegvis regressionsanalys. Jag lämnar här de tekniska detal-

jerna och går direkt till ett par exempel som sammanfattar allt det viktigaste i analysen och verkligen ger skäl för beteckningen longitudinell i rubriken.

EXEMPEL PÅ STIGANALYSER

Tabell 6. Mäns användning av »positiva» ordklasser.

	<i>Språk 13</i>	<i>Utbildning</i>	<i>Språk 32</i>	<i>Andel i "positiva" ordklasser</i>
<i>Socialgrupp</i>	0.14	0.14	(-0.04)	(0.05)
<i>Språk 13</i>		0.61	0.23	0.19
<i>Utbildning</i>			0.33	0.27
<i>Språk 32</i>				0.20

Multipl korrelation 0.56

För männen har de förklarande variablerna socialgrupp till och med Språk 3 tillsammans en multipel korrelation med andelen »positiva» ordklasser på 0.56, vilket är ett ganska högt värde. Socialgrupp påverkar i måttlig grad språk vid 13 års ålder och utbildningsnivå. Den har däremot inga direkta effekter på läget vid 32 och 34 års ålder. Språkförmågan enligt våra mätningar vid 13 års ålder påverkar starkt den utbildningsnivå som individerna nått vid 32 års ålder och likaså fast i svagare grad ordanvändningen i intervjun. Den uppnådda utbildningsnivån har en hygglig effekt på prestationerna hos de vuxna. Det mest avgörande steget är emellertid från språk 13 via utbildning till vuxet språk.

Tabell 7. Kvinnors användning av »positiva» ordklasser

	<i>Språk 13</i>	<i>Utbildning</i>	<i>Språk 32</i>	<i>Andel i "positiva" ordklasser</i>
<i>Socialgrupp</i>	(0.07)	0.15	0.11	(0.00)
<i>Språk 13</i>		0.58	0.26	(0.01)
<i>Utbildning</i>			0.34	0.30
<i>Språk 32</i>				0.35

Multipl korrelation 0.59

För kvinnorna är bilden någorlunda likartad bortsett från att Språk 13 inte har någon direkt effekt alls på andelen »positiva» ordklasser, utan hela sambandet till vuxen ålder går via utbildningsnivån. Den multipla korrelationen är 0.59, gott och väl så stark som hos männen.

Sådana här analyser har gjorts också för alla enstaka ordförrädsvariabler. Den variabel bland alla som har den högsta multipla korrelationen visade sig vara genomsnittslängden på de ord som använts bara en gång i intervjun.

Detta mått sammanför två kvalitetsaspekter; sällsynta ord och komplexa ord. Där var multipelkorrelationen för män 0.64 och för kvinnor 0.58 med utbildningsnivå som främsta bidragande variabel.

Med detta slutar jag denna lilla provkarta på vad analyser av intervjuspråk kan ge när de görs i relation till tidiga mått på individernas språkliga nivå och utbildningens roll i sammanhanget. Analyserna kan ses som ett sent bidrag till vårt 1980-talsprojekt om »långtidseffekter av utbildning».

LITTERATUR

- Allwood, J. (red) 1999: *Talspråksfrekvenser: frekvenser för ord och kollokationer i svenskt tal- och skriftspråk: frekvenser för ord, ordklasser och kollokationer: jämförelser mellan tal och skrift*. (Gothenburg Papers in Theoretical Linguistics) Göteborg: Göteborgs universitet, Institutionen för lingvistik.
- Härnqvist, K., Christiansson, U. & Tingsell, J-G. 1985: *LING-projektets intervju. Kvantitativa mått på språkanvändning och deras samband med utbildningsnivå och kön*. Göteborg: Göteborgs universitet, Pedagogiska institutionen.
- Härnqvist, K., Christiansson, U., Ridings, D. & Tingsell, J-G. 2003: Vocabulary in interviews as related to respondent characteristics. *Computers and the Humanities*, 37(2), 179–204.