

Fönster mot språkvärlden.

Om användningen av textdatabaser (korpora) i undervisning, forskning och textproduktion

Av Maria Estling, doktorand i engelska

[Länk till presentation av Maria Estling](#)

- [Vad är en korpus?](#)
- [Korpora i forskning](#)
- [Korpora i undervisning](#)
- [Korpora i textproduktion](#)
- [Nyhetsbrevet GramTime News](#)

Vad är en korpus?

Den som kan sitt latin vet att *corpus* är det latinska ordet för kropp. Inom språkforskningen används ordet för att beskriva en samling texter, och i dagens IT-samhälle används det ofta synonymt med en datoriserad textsamling, lagrad antingen i en hårddisk eller på en CD-ROM-skiva.

Det man i första hand förknippar med ordet *text* är förstas en skriven sådan, t.ex. en novell eller en tidningsartikel. I det här sammanhanget har dock ordet en avsevärt vidare betydelse, och används om såväl talat som skrivet språk. En text kan således vara en föreläsning på universitetet, ett telefonsamtal eller ett samtal på ett kafé. Språkforskare som sysslar med talad text använder sig vanligen av inspelningar som *transkriberats* (skrivits ner) och som därigenom blir lättare att hantera, men man kan också arbeta direkt med inspelningarna.

Eftersom mitt forskningsområde är engelska kommer de exempel jag ska visa från den engelsktalande världen. Det är också ett faktum att en mycket stor del av de datoriserade korpora som används idag är uppbyggda av texter på engelska, och att den korpusforskning som hittills gjorts i stor utsträckning gällt det engelska språket.

Textmaterialet kan bearbetas på olika sätt, men vanligast är att det används tillsammans med ett sökprogram som t.ex. kan skapa så kallade *konkordanslistor* i stil med följande:

```
joint, with a beautiful woman shimmering opposit  
take in the complete woman at one swallow. Th  
she was a very capable woman. <p> With a few Sh  
was a once-handsome woman in her fifties, wh  
I knew she was a sensible woman. It was what I had  
wouldn't vote for the silly woman except that there  
something. That silly woman put it out of my  
as the confident, capable woman who had come to Ye  
Mrs McNiece?" <p> A young woman, looking hot and  
to do so. She was a tall woman thirty-eight was a  
acute, intelligent woman. <p> But you didn'  
of course." <p> Wonderful woman," Henry said  
sincere. She was a small woman, very slender, ver  
as a small, plump elderly woman in a dress of the  
head of a beautiful woman on one side and a  
solid, cheerful, untidy woman with her grey hair  
verything - and that damned woman." She pulled
```

I mitten ser vi sökordet - *woman* (föregått av ett adjektiv) i det här fallet - och med varje

förekomst får vi också en bit omgivande text. Den databas som använts för att få fram den här listan har ett av de mest användarvänliga sökprogrammen, och man kan med några enkla kommandon bland annat:

- ta reda på från vilken text varje förekomst är hämtad
- få fram mer omgivande text
- sortera orden före eller efter sökordet (inom valfritt avstånd, t.ex. två ord till vänster om sökordet)
- se vilka ord som är vanligast i sökordets omgivning (inom valfritt avstånd)
- spara ner listan till ett Word-dokument så att man kan redigera och använda den som man vill (t.ex. i en text som jag har gjort här)

På detta vis kan man söka information om hur det engelska språket används i olika delar av världen, och jag ska under de tre följande rubrikerna ge exempel på hur denna information kan vara till nytta inom olika områden.

Korpusar i forskning

Inom forskarvärlden är användning av korpusar inte något nytt fenomen. Språk- och litteraturforskare har i många år skapat sina egna textsamlingar och förvarat dem på lappar eller arkivkort. När jag skrev min D-uppsats i engelska för ett antal år sedan byggde jag - utan att känna till ordet - upp en korpus av svordomar ur några romaner, vilken jag sedan använde för att undersöka hur svordomsbruket kan se ut i amerikansk engelska. Något som är betydligt nyare är forskning med hjälp av datoriserade korpusar, vilken startade under sextiotalet. På den tiden var korpusarna jämförelsevis små (500.000 - 1.000.000 ord), och datamaskinerna var dyra, svårhanterliga och långsamma. Nittiotalets IT-revolution har visat sig även inom korpusforskningen. Kraftfulla datorer gör det numera möjligt att skapa textdatabaser med hundratals miljoner ord, och sökningar i dessa går snabbt och är oftast enkla att utföra.

Inom forskningsprojektet *GramTime (Grammatical Trends in Modern English)* vid Växjö universitet använder vi ett antal mycket stora textdatabaser med olika typer av talad och skriven text. Dessa korpusar består bl.a. av hela tidningsupplagor, romaner, faktaböcker, vetenskapliga artiklar, radioprogram, telefonsamtal och vardagskonversation. Syftet med vårt projekt är att kartlägga dialektal och stilistisk variation och förändring i modernt engelskt språkbruk (framförallt inom grammatiken). Vi undersöker skillnader mellan brittisk, amerikansk, australiensisk och nyzeeländsk engelska, mellan talat och skrivet språk och mellan olika textgenrer. Denna information kan sedan användas för att modernisera och förbättra läromedel och referensverk. För att sprida våra forskningsresultat ger vi också ut ett nyhetsbrev, *GramTime News* - mer om detta längre fram. Språkvariation och språkförändring är populära områden inom korpusforskningen och över hela världen finns forskare som sysslar med liknande saker.

Ett annat viktigt område inom korpusforskningen är det pedagogiska. Det finns en speciell typ av korpusar, så kallade *inlärarkorpusar*, som är uppbyggda av studentuppsatser författade av språkstudenter av olika nationaliteter, bl.a. svenska. Genom att jämföra detta textmaterial med uppsatser skrivna av infödda engelsktalande - samt med korpusar med mer blandat textmaterial - kan man få fram information som är av stor nytta för språklärare och författare av undervisningsmaterial. Man kan t.ex. göra felanalyser för att ta reda på vilka typiska fel icke infödda gör, och man kan se vilka konstruktioner eller ord som icke infödda överanvänder eller underanvänder i jämförelse med infödda engelsktalande personer. På detta sätt kan man komma fram till vad man bör fokusera på i språkundervisningen. Korpusarna är också en utmärkt källa till exempel i kursböcker. Alla som studerat språk har stött på de traditionella grammatikböckernas konstruerade, ofta onaturliga exempel som ibland till och med har hängt med från en bok till en annan. Nog känns det betydligt mer stimulerande att lära sig ett språk när bokens exempel är hämtade från autentisk text.

Andra korpusområden är bl.a.

- forskning om ett språks utveckling genom historien (med hjälp av korpusar med material från tidigare språkperioder, t.ex. den engelska som talades under medeltiden eller under 1960-talet)
- ordspråksforskning (identifikation av ett uppslagsords betydelse och vilken/vilka betydelse(r) av ordet som är vanligast)
- forskning om hur nya ord uppkommer och etableras i eller försvinner från språket
- jämförelser mellan manligt och kvinnligt språk

En gång om året möts ett stort antal korpusforskare till en konferens där nya korpusar, nya sökprogram och resultat av korpusforskning presenteras och diskuteras.

Korpusar i undervisning

Som lärare i engelska ställs man ofta inför frågor som är svåra att finna svar på i grammatikböcker och lexikon. Jag har själv stor nytta av mina korpusar när studenterna jag undervisar i engelsk grammatik kommer med funderingar i stil med följande:

- "Kan man uttrycka sig såhär? Det står inte i grammatikboken, men jag hörde folk säga så när jag pluggade utomlands?"
- "Vilken konstruktion är vanligast, den här eller den här?"
- "Finns den här konstruktionen bara i brittisk (eller amerikansk) engelska?"

När man rättar uppsatser ställs man också ofta inför konstruktioner som man inte känner igen och kanske inte hittar i böckerna. Genom att titta i en eller ett par korpusar kan man ofta få en fingervisning om en konstruktion är vanlig, ovanlig eller inte förekommer alls. Därmed inte sagt att korpusar ger hela sanningen - det är ju inte Svea Rikes lag det är fråga om, bara en samling autentiska texter. Detta är viktigt att ha i åtanke när man uttalar sig om företeelser utifrån korpusen. Man kan t.ex. naturligtvis inte från en enda förekomst av ett ord eller uttryck dra slutsatsen att detta är korrekt engelska. På samma sätt måste inte avsaknaden av ett uttryck nödvändigtvis betyda att det inte existerar i språket. Ibland blir man dock förvånad över att finna att konstruktioner som inte alls är tillåtna enligt grammatikböckerna faktiskt är vanligt förekommande.

Vid Växjö universitet låter vi också studenter i engelska på C- och D-nivå själva arbeta med korpusar när de skriver sina uppsatser. Tillgången till ett omfattande autentiskt textmaterial är en stor fördel när man ska behandla ett språkvetenskapligt problem. Att arbeta med datorkorpusar går naturligtvis mycket snabbare och materialet blir mer allmängiltigt än om man t.ex., som jag själv gjorde, skapar en egen liten korpus med bara några romaner som underlag och ska hålla reda på alla exempel för hand.

Även inom den vanliga skolan (åtminstone på gymnasienivå) kan man arbeta med korpusar, t.ex. i form av en CD-ROM-skiva med tidningstext där eleverna själva kan utforska det språk de ska lära sig. Man kan också använda Internet som en enda stor korpus för att söka på olika ord och uttryck i ett språk. Här måste man förstås vara försiktig om man vill uttala sig om t.ex. engelskan så som den talas av infödda, eftersom en stor del av de texter på engelska man hittar på nätet är skrivna av icke infödda talare.

Korpusar i textproduktion

Den som använder ett annat språk än sitt modersmål har stor nytta av att ha tillgång till en korpus. Vid direktproduktion av texter på ett främmande språk eller översättning från modersmålet kan korpusen t.ex. ge hjälp att hitta *idiomatiska uttryck*, d.v.s. för ett språk typiska och ej direkt översättningsbara konstruktioner, i målspråket. Sådana konstruktioner är ofta svåra att hitta i ordböcker. Ett exempel på ett idiomatiskt uttryck som vi inom

projektet förgäves letade efter i våra lexikon är den engelska motsvarigheten till den svenska frasen *som att jämföra äpplen och päron*, med betydelsen 'att jämföra två ting eller företeelser som egentligen inte är jämförbara'. Vi sökte på *apples* i våra korpusar och hittade på så sätt uttrycket: *to compare apples and oranges*. Här följer några exempel ur textmaterialet:

We can't have five different measurement systems out there, Ms. Black said, because it is vital to compare **apples with apples and oranges with oranges**.

They are available, of course, but comparing **apples** (Germany) with **oranges** (the rest) doesn't help reach any conclusions.

"People don't treat China shares like the **apples and oranges** that they are," said Mr Tim Greaton, who manages \$120 million in China funds at Credit Lyonnais.

Det är också betydligt lättare att på det här viset hitta den mest lämpliga konstruktionen i ett speciellt sammanhang då det finns flera olika sätt att uttrycka sig på.

Vidare kan man få hjälp när man översätter till sitt eget modersmål. Ett uttryck i ursprungstexten som inte finns beskrivet i lexikon, och inte går att tolka utifrån den omgivande texten, kan ibland få sin förklaring genom att man hittar fler exempel i korpusen.

Nyhetsbrevet **GramTime News**

Avslutningsvis skulle jag vilja säga några ord om det nyhetsbrev som nämndes tidigare. Vi som arbetar inom *GramTime*-projektet i Växjö tycker att det är viktigt att den information vi får fram i vår forskning inte stannar inom en liten krets av andra forskare, utan också når ut till världen utanför. Av denna anledning startades våren 1998 ett nyhetsbrev, *GramTime News*, som främst riktar sig till lärare runt om i Sverige och världen, men också till andra språkintresserade personer. Nyhetsbrevet kommer ut 4-5 gånger per år i elektronisk form, d.v.s. brevet skickas ut som ett e-postmeddelande och publiceras samtidigt på vår hemsida på Internet, där man kan läsa texten i ett mer behagligt format. *GramTime News* är gratis och öppet för alla. I nyhetsbrevet försöker vi med hjälp av våra korpusar besvara frågor om modernt engelskt språkbruk från våra läsare, sådant som man som språklärare eller språkstudierande kanske undrat över men inte hittat svar på i böckerna. Här följer några exempel på vad *GramTime News* har behandlat hittills. Svaren är förkortade, och mer utförlig information kan man få genom att titta på gamla nummer på hemsidan (se nedan).

- Kan förkortningar som *USA* och *EU* användas utan bestämd artikel?

Svar: Infödda talare verkar bara använda förkortningarna utan artikel i rubriker (*EU agrees truce in Eta extradition*) och uppräknningar (*Iraq severs diplomatic relations with Egypt, France, Italy, Saudi Arabia, UK and USA*).

- Används ord som *scissors*, *police* och *pyjamas* någonsin i singular form, t.ex. *Has the police caught the murderer yet?*

Svar: Ja, det förekommer vid vissa ord (t.ex. *scissors* och *police*), men är inte särskilt vanligt.

- Är mängduttrycket *plenty of* begränsat till engelskt talspråk?

Svar: *Plenty of* var mycket vanligt i alla typer av text, även om det fortfarande förekommer mest i informell engelska.

Kan man använda *Such* i stället för *What* i utrop, som i *Such/What a nice girl!*

Svar: *Such* (som ofta rättas till *What* av nitiska engelsklärare) förekom ganska ofta i de texter vi undersökte, även om *What* var betydligt vanligare - det användes i ca 90% av fallen.

- Vilken är den vanligaste översättningen av 'varannan', *every second*, *every other* eller *every two*?

Svar: *Every two* var betydligt vanligare än de andra uttrycken. *Every other* är vanligare i amerikansk engelska än i brittiskan. *Every second* kan precis som *every other* användas för att uttrycka en mindre precis uppräkningsord än 'varannan', ofta med en irriterad underton ('var och varannan').

- Kan man använda *a great deal of* tillsammans med substantiv i plural, t.ex. *a great deal of students*?

Svar: I svenskan används ju uttrycket *en hel del* ofta tillsammans med substantiv i plural (t.ex. *en hel del människor*). I engelskan är däremot *a great deal of* mycket ovanligt vid ord i plural.

- Förekommer uttryck som *out the window* (som brukar betecknas som en amerikansk variant av det brittiska *out of the window*) i brittisk engelska?

Svar: Ja, det är till och med det vanligaste uttryckssättet i talad brittisk engelska (60-70%), och även ganska vanligt i skriven brittisk engelska, särskilt i tidningstext. Det "brittiska" uttrycket *out of* finns också i amerikansk engelska.

Förutom språkbruksfrågor innehåller *GramTime News* också bokrecensioner och tips om användbara websidor. Information om prenumeration m.m. får man på <http://www.vxu.se/hum/publ/gtn/> eller genom att kontakta projektet på e-postadress gramtime@hum.vxu.se.

© [Maria Estling](#)