Sture Allén

INTE BARA IDIOM

Detta skall handla om kollokationer eller ordförbindelser. Vissa grundfrågor inom språkvetenskapen upphör inte att fängsla en, och kollokationerna har för mig varit en sådan grundfråga. Bland mycket annat har de ett intressant förhållande till parsning.

Jag minns med stor intensitet mina intryck, när jag såg de första resultaten av konkordanskörningar för ganska många år sedan. De öppnade nya språkliga vyer. Man såg i ett huj vilken betydelse kollokationerna måste ha i den språkliga aktiviteten. Detta har varit en viktig utgångspunkt för ständigt uppdaterade tankar kring kollokationer.

En av de första jag stötte på som över huvud taget hade tänkt i de här banorna med datamaskinen i perspektivet var John Sinclair (1970). Han menade att man skulle ta fasta på vad som är statistiskt signifikant för att få fram kollokationerna. De ord som uppträdde tillsammans oftare än man hade anledning att förvänta sig, sett mot en slumpmässig bakgrund, skulle ha kollokationell karaktär. Kriteriet gav både relevanta och irrelevanta (*if take* osv.) förbindelser. Resultatet av arbetet var således inte övertygande i detta avseende, inte heller för Sinclair själv. Också andra experiment har sedan visat att det naturligtvis inte är så enkelt, även om det är en del av sanningen.

De första tankarna med utgångspunkt i en allmän uppfattning om språket och om vad konkordanserna visade ledde till uppläggningen av arbetet på tredje delen av Nusvensk frekvensordbok (1975) och presentationen av resultaten. Jag talade under 1970-talets första hälft bland annat i Åbo, Pisa och London om kollokationerna och deras roll i språket (föredragen trycktes 1973, 1976 respektive 1977).

Det som var utgångspunkten för tredje delen av frekvensordboken var iakttagelsen att kollokationerna ständigt återkommer,

Inte bara idiom
Sture Allén
Proceedings of NODALIDA 1983, pages 14-20

är rekurrenta i den mening som anges i inledningen till frekvensordboken. Villkoret i denna undersökning var att det skulle finnas minst två identiska belägg på en miljon löpande ord. Med det villkoret fick vi ut 660 000 belägg på ordkombinationer. Det var kanske den första överväldigande siffran för oss. Den vanligaste av alla förbindelserna var *det är*, sedan kom *och den*, *sig i*, *än att*, *på ett helt annat sätt* osv., alltså välformade och icke välformade ordkombinationer om vartannat.

Ur de 660 000 beläggen analyserade vi fram vad vi kallade för konstruktioner. På dem lade vi villkoren att de skulle vara grammatiskt styrda och lexikaliskt selekterade. De utgör sålunda ett urval ur kombinationerna. Vi fick 50 000 olika konstruktioner, som vi klassade i 17 olika huvudtyper: *den stora frågan*, *i linje med*, *ta form*, *sköta om*, *kommer att fortsätta*, *för att undersöka*, *är på väg*, *mycket ung*, *även om*, *om också*, *för att*, *in i*, *men låt oss inte gå händelserna i förväg*, *det är givet att*, *kort sagt*, *som människa*, *jo då*. Dessa exempel markerar de olika typerna.

Ur konstruktionerna gjorde vi i sin tur ett urval. Det gällde idiomen i ordets snäva bemärkelse, dvs. de kollokationer som har oförutsägbar betydelse sett från de ingående ordens betydelsers synpunkt. Vi tillämpade kriterierna ganska strängt och fick fram 300 olika idiom: *ge sig i kast med*, *i elfte timmen*, *lägga sista handen vid*, *det är inte utan att*, *göra avkall på*, *slå dövörat till* osv. Idiomen utgör alltså en mycket liten del av den stora mängden kollokationer.

I den nämnda inledningen påpekar jag, att det är ganska tydligt att resultaten bör få konsekvenser för språkteorin. Lexikonet ser antagligen inte ut på det sätt som man har tänkt sig tidigare, t. ex. inom den generativa inriktningen. Vad man kan kalla lexikaliska block av många olika slag bör finnas i språkmedvetandet.

Strax efter utgivningen av frekvensordbokens tredje del publicerade Jan Anward och Per Linell (1976) en uppsats om lexikaliserade fraser i svenskan. De tog fonologi, grammatiska egenskaper m.m. i betraktande och sammanfattade sitt resonemang så

här: varje enskild konstituent i en lexikalisk enhet (dvs. i en kollokation) kan inte exploatera sin betydelse fullt ut och kan inte syntaktiskt varieras i någon större utsträckning; kan inte ha egen referens (säger man *sparka boll*, syftar man inte på en speciell boll); kan inte böjas fritt; kan inte modifieras fritt; kan inte fritt befrågas, negeras eller affirmeras; kan inte varieras med avseende på ordföljd och prosodi i någon större utsträckning; kan å andra sidan utmärkas av morfologiska och andra oregelbundenheter.

Om man tittar lite närmare på detta, ser man att flera av villkoren är av sådan karaktär att de karakteriserar vad jag vill kalla idiom snarare än kollokationer i allmänhet. Men slutsatsen är i alla fall densamma som den jag hade dragit från andra utgångspunkter, nämligen att lexikonet spelar en större roll än man kanske trott och att grammatiken spelar en något mindre roll.

Charles Ruhl och Adam Makkai förde en diskussion om idiom i en volym från 1976. Utgångspunkten var att Makkai hade skrivit en bok om Idiom structure in English. Ruhl menade att ett fel med den boken var att Makkai inte hade tagit hänsyn till den omedvetna delen av språkmanifestationen. Han tog som exempel *take off* som ju på engelska betyder 'become airborne' men som i sin nominaliserade form också kan betyda 'parody'. Makkai menade att man här har två olika lexikaliska enheter medan Ruhl efter en genomgång av ett hundratal belägg på *take off* noterar hur man från den ena betydelsen gradvis kommer över i den andra.

På detta svarar Makkai genom att ta fram alla ord som i engelskan slutar på *-ic(k)* (*bic*, *brick*, *chick*, *dick* osv.) och tillsammans med sina studenter resonera sig fram till hur vart och ett kan anknytas till dels 'liv' eller 'död', dels 'skatter'. Med andra ord: från vad som helst kan man resonera sig fram till vad som helst. Det är ett sätt att ironiskt avfärda den tanke som Ruhl hade. Man måste försöka identifiera de olika lexikaliska enheterna, menar Makkai. Han använder rentav benämningen *institutions* för vissa av de här kollokationerna – de har så att säga en stadfäst roll i språket.

Ungefär samtidigt hade vi i Göteborg ett mindre projekt som hette Algoritmisk textanalys. Jag förbigår här det som gjordes på programsidan och i någon mån på syntaxsidan och nämner de lexikaliska delresultaten. Vi utarbetade ett paradigmmärkt baslexikon på ungefär 8000 enheter. Det är publicerat av Staffan Hellberg i en bok från 1978. Baslexikonet upptar stam, uppslagsform och paradigmnummer och täcker därmed i princip hela morfologin. Vidare upprättades ett speciallexikon över heterografer (ickehomografer) på ungefär 900 enheter. Ett tredje resultat var ett speciallexikon över disambiguerande kollokationer på omkring 1600 enheter. Det omfattar sådana kollokationer som innehåller homografer som blir disambiguerade genom att ingå i kollokationerna. Ett exempel är *komma hem*, där *komma* kan vara verb eller substantiv och *hem* kan vara substantiv eller adverb, medan *komma hem* är entydigt. Till detta kom ett ändelselexikon för sannolikhetsklassificering av ord som saknas i baslexikonet. Ett av de ständigt återkommande problemen är ju att man stöter på nya ord.

Bland det mest intressanta var att de två speciallexikonen på 1600 respektive 900 enheter vid körningar visade sig täcka 50 procent av en okänd text av normal typ. Det är tankeväckande från parsningens synpunkt. Jag kan tillägga att de 8000 enheterna i baslexikonet täcker i runda tal 90 % av en text. Då skall vi emellertid komma ihåg att de innefattar en mängd homografer. Det fina med de två speciallexikonen är att de ger stycken av fast mark som analysen kan bygga vidare på.

Låt mig nämna ytterligare några som på olika sätt har arbetat med kollokationer. En av dem är Harald Burger som har publicerat en intressant bok om idiom (1973), där han framför allt är inne på de teoretiska problem som knyter sig till begreppet. Maurice Gross (1982) har undersökt vad han kallar "frozen sentences", vilket också är ett slags kollokationsbegrepp. I hans fall har det gällt franska. Göran Kjellmer vid engelska institutionen i Göteborg är sysselsatt med en genomgång av hela Brown-korpusen för att ta fram kollokationsmaterialet ur den. Syftet är främst att utarbeta en frasordbok.

Så är vi framme vid Lexikalisk databas. Det är det största projektet vid Språkdata för närvarande. I det definierar vi omkring 75 000 lemman ur det moderna svenska språket och ger uppgifter av många olika slag, bl.a. just beträffande fraseologi och idiomatik.

Här är ett utdrag ur kollokationsuppgifterna rörande lemmat *land*, som representerar tre olika lexem (lexikaliska enheter baserade på kärnbetydelser). Jag ger inte deras definitioner, utan vi kan se på exemplen vad de avser. Till det första lexemet hör *inom landets gränser, flytta till ett annat land, de afrikanska länderna* och idiomen *det heliga landet* eller *det förlovade landet* och *skuggornas land*. Det andra omfattar *en sjöman går i land, land i sikte, på torra land, färdas till lands* och idiomen *förstå hur landet ligger, gå i land med något* och *nu går skam på torra land*. Det tredje har *hon var från landet, resa till landet under veckoslutet* och idiomet *ingen mans land*. En av tankarna med projektet Lexikalisk databas är just att man med utgångspunkt från i första hand kollokationerna och definitionerna skall arbeta vidare i riktning mot ett lexikon för parsning. I existerande system är de såvitt bekant, liksom lexikaliska uppgifter över huvud, svagt företrädda. Låt mig ändå i detta sammanhang nämna Kaplan & Bresnan (1980), Sager (1981) och Zimmermann, Kroupa & Keil (1983).

På senare tid har jag kommit att uppmärksamma en annan typ av indikationer som jag anser vara viktig från såväl teoretisk som praktisk och psykologisk synpunkt. Det gäller vad jag kallar för de metaspråkliga kommentarerna i texter. Det är alltså så att språkbrukarna själva i viss utsträckning talar om hur deras lexikon ser ut, något som man kanske inte har uppmärksammat tidigare. De metaspråkliga kommentarerna gäller rätt ofta enskilda ord men inte sällan just kollokationer. Man markerar dem med uttryck av typen *som det heter, som det så vackert heter* osv. Ett litet urval exempel följer: *ett rörligt intellekt, som det heter; karavanen rör sig trots att hundarna skäller, som det heter i ett gammalt arabiskt ordstäv; administrativ databehandling, som det heter; ett förslag till, som det så vackert heter, en förenklad deklarationsblankett; bidrag för att,*

som det heter, förbättra konkurrensmöjligheterna på den internationella marknaden; hon är väl död vid det här laget, om man så säger; vi nådde alltså, för att lätt travestera ett slitet uttryck, ända fram.

Själva de metaspråkliga uttrycken är rikt varierade och baserar sig på en rad olika verb: *för att använda en kliché, för att använda ett gammalt ordspråk, för att använda ett slitet uttryck, för att använda herr NNs egen formulering; för att citera NN; som det heter, som det numera heter, som det brukar heta, som det så vackert heter; som man brukar kalla det, som det kallas, så kallad; som frasen lyder, som uttrycket lyder; som man säger, så att säga, om jag så säger, som det brukar sägas; för att travestera ett gammalt uttryck; om man så vill; om uttrycket tillåts, om man så får uttrycka saken, som NN uttrycker det.* Alla de metaspråkliga kommentarerna syftar inte alltid på kollokationer i den mening jag avser här, men mycket ofta visar det sig vara fallet. Man kan dela in dem i några huvudtyper. *Som det heter* syftar ofta på en kurant kollokation. *Som X lyder* syftar gärna på ett ordspråk eller ordstäv. *För att citera* anger direkt källan: en författare, en lagtext, en förordning eller någonting sådant. *För att travestera* kan i princip syfta på alla de olika typerna. Som travesti åberopar den indirekt en kollokation av något slag.

De metaspråkliga kommentarer som man kan plocka fram på det här sättet ger ytterligare en inblick i människans lexikon. De fogar sig också till de tidigare typerna av kollokationella kriterier som har varit rekurrensen, de grammatiska konstruktionskriterierna, idiomkriteriet och de tillkommande lingvistiska kriterier som Anward och Linell har pekat på. De bidrar alltså till att ge en ny bild av lexikonet och följaktligen också en ny bild av hur vi fungerar i språkproduktionen och i perceptionen. Därmed är de av fundamentalt intresse vid utvecklingen av parsningsystem.

Litteratur

Allén, Sture, Om fraser i svenskan. (Svenskans beskrivning 7. Ed. Christer Hummelstedt. Åbo 1973, s. 24-31.)

Allén, Sture, On phraseology in lexicology. (Cahiers de lexicologie 29 (1976), s. 83-90.)

Allén, Sture, Text-based lexicography and algorithmic text analysis. (ALLC Bulletin 5: 2 (1977), s. 126-131.)

Allén, Sture, et al., Nusvensk frekvensordbok baserad på tidningstext. 3. Ordförbindelser. 1975.

Anward, Jan, & Linell, Per, Om lexikaliserade fraser i svenskan. (Nysvenska Studier 55-56 (1975-76), s. 77-119.)

Burger, Harald, Idiomatik des Deutschen. Tübingen 1973. (Germanistische Arbeitshefte. Ed. Otmar Werner & Franz Hundsnurscher. 16.)

Gross, Maurice, Simple sentences. (Text processing. Proceedings of Nobel Symposium 51. Ed. Sture Allén. 1982, s. 297-315.)

Hellberg, Staffan, The morphology of Present-Day Swedish. Word-inflection, word-formation, basic dictionary. 1978.

Kaplan, R.M., & Bresnan, J.W., Lexical-functional grammar: a formal system for grammatical representation. Occasional Paper 13. MIT Center for Cognitive Science. Cambridge, Mass. 1980.

Makkai, Adam, Idioms, psychology, and the lexemic principle. (The Third Lacus Forum. Ed. Robert J. Di Pietro & Edward L. Blansitt, Jr. Columbia, South Carol. 1976, s. 467-478.)

Ruhl, Charles, Idioms and data. (The Third Lacus Forum. Ed. Robert J. Di Pietro & Edward L. Blansitt, Jr. Columbia, South Carol. 1976, s. 456-466.)

Sager, Naomi, Natural language information processing: a computer grammar of English and its applications. Reading, MA 1981.

Sinclair, J. McH., Jones, S., & Daley, R., English lexical studies. Department of English. Birmingham 1970.

Zimmermann, Harald H., Kroupa, Edith, & Keil, Gerald, CTX. Ein Verfahren zur computergestützten Texterschliessung. Saarbrücken 1983.