GÖTEBORGS UNIVERSITET
SPRÅKDATA
Bo Ralph
November 1979

LEXIKOLOGI SOM DATALINGVISTIK

Alla teoretiska språkmodeller brukar ha en komponent med språkliga byggstenar, ett lexikon, och en komponent med regler för byggstenarnas sammanfogning till satser och texter, en grammatik. Vid lemmatisering – angett som ett huvudtema vid datalingvistikdagarna – aktualiseras ett par grundläggande frågor som rör den lexikaliska komponenten. Med lemmatisering syftar man i första hand till att samklassificera formella och funktionella varianter till abstraktare lexikaliska enheter och upprätta kanoniska grundformer. Jag vill med det här bidraget lägga några generella synpunkter på lexikaliska enheters egenskaper och inbördes relationer, dvs. på den lexikaliska komponentens allmänna struktur.

Det lexikaliska modellbyggandet är i och för sig inte någon specifikt datalingvistisk angelägenhet. Det finns emellertid flera skäl att diskutera lexikologi i ett datalingvistiskt sammanhang. För det första har inte allmänlingvisterna ägnat sig i särskilt hög grad åt lexikaliska frågor. Huvudvikten har alltid legat vid andra grammatikkomponenter, såsom den syntaktiska, den semantiska, den fonologiska och i någon mån den morfologiska, olika starkt betonade under olika perioder. Eftersom den teoretiska modellen ändå förutsätter en lexikalisk komponent har denna tenderat att bli något slags "garbage component", som man hänskjutit problemen till då man inte velat ta itu med dem i det sammanhang som just varit aktuellt. Som kontrast har datalingvister ofta haft att lösa lexikaliska problem inom ramen för praktisk verksamhet. Datalingvisten har ofta nog fått bli sin egen lexikolog. Som en följd har rätt mycket av den teoretiskt inriktade lexikologi som över huvud taget bedrivits på senare tid presterats av datalingvister.

Det finns emellertid skäl även för andra lexikologer att närma sig datalingvistiken. Själva det lexikaliska materialets art gör det nödvändigt att lexikologi bedrivs i nära samråd med data-

Lexikologi som datalingvistik
Bo Ralph
Proceedings of NODALIDA 1979, pages 161-170

lingvister. Välkänd är Dr. Johnson's karakteristik av lexikografen som "a harmless drudge that busies himself in tracing the original and detailing the signification of words". Även lexikologi är i långa stycken ett träljobb; i högre grad än inom de flesta andra lingvistiska forskningsgrenar förutsätts det att stora datamängder undersöks. Datorn är den idealiske trälen. Metodologiskt ställs lexikologin i ett helt nytt perspektiv genom möjligheterna till interaktiv bearbetning. Struktureringen accentueras om fasta format används. Modellen vinner stadga genom möjligheterna till smidig kategorivis genomgång. Men framför allt är själva den experimentella inriktning som präglar datalingvistiken en tillgång. Trycket på många datalingvister att komma till praktiska resultat ger upphov till idéer som bör prövas inom en allmänteoretisk ram. De systematiseringar av datamängder som utförs av informationsbehandlare och de informationsstrukturer som därigenom byggs upp kan likaledes skänka inspiration åt lexikologin. En verksamhetsgren som i synnerhet måste bevakas av lexikologer är databastekniken, som säkerligen kan ge impulser till lexikologiskt nytänkande.

När lexikon skapas i datalingvistiska sammanhang sker det kanske vanligen med siktet inställt på att klara av en given uppgift. Det är naturligt och legitimt att enheterna i diverse maskinlexikon kan ha högst varierande karaktär beroende på syftet. Avvägningen mellan lexikonets och grammatikens respektive kraftfullhet, fördelningen mellan rent automatisk och interaktiv bearbetning etc. blir praktiskt baserade optimeringsproblem. Avser man att approximera människans eget sätt att fungera språkligt – och det kan man naturligtvis göra även om det primärt är ett praktiskt problem som skall lösas – ställer sig saken annorlunda. Då kommer t.ex. resonemanget om psykologisk relevans in i bilden. Hela synen på den lexikaliska enheten påverkas om hänsyn tas till den psykologiska verkligheten.

Lingvister tycks ofta föreställa sig det mänskliga lexikonet som något slags idealiserad ordbok av relativt traditionell utformning. Det betyder att man övertar åtskilliga av de problem som förknippas med ordböckers utformande. Det är exempelvis en grundläggande fråga vad som skall konstituera en lexikalisk enhet. Problemet har sin rot i förhållandet mellan form och betydelse.

Så länge relationen är någorlunda ett-till-ett, som i *distikon*, *hata*, *vit*, är läget okomplicerat, men hur många lexikaliska enheter bildar det berömda *krona*? Det kan betyda 'huvudprydnad för furste', 'konungamakten el. staten som institution', 'den övre delen på träd', 'den greniga delen på hjortdjurs horn', 'utstyrd taklampa' och 'myntenhet i Norden m.fl. områden' osv. Sture Allén har för svenskans del diskuterat dessa problem i flera skrifter (t.ex. i inledn. till NFO 1 och i Allén 1967).

I den livliga lingvistiska debatten under sextiotalet tilldrog sig knappast den lexikaliska komponenten något huvudintresse, men några utkast till lexikonstruktur som då presenterades är typiska för tiden. Katz och Fodor (1963) tänkte sig att en lexikalisk enhet kunde ha underordnade betydelsevarianter (*bachelor*), ungefär som i traditionell lexikografi. McCawley (1968) uppfattade varje ny kombination av ett uttryck och ett innehåll som en egen lexikalisk enhet. Dessa båda infallsvinklar ger som resultat höggradig polysemi respektive homonymi. I övrigt har diskussionen kring den lexikaliska enheten mera gällt vilken information som skall knytas till denna än hur själva enheten skall avgränsas.

I traditionell lexikografi företas uppdelningen som bekant på etymologins grund. Därigenom blir *tunga* 'slags plattfisk' och *tunga* 'rörlig muskelkropp i munhålan' betraktade som e t t uppslagsord med polysemkomponenter, medan *axel* 'skuldra' är en självständig lexikalisk enhet helt skild från *axel* 'tvärstång el. tänkt centrum kring vilken/vilket något roterar' (jfr fsv. *axl* resp. *axul*). Från strikt synkronisk synpunkt är det inte mer naturligt att associera fisken till muskelkroppen än att associera skuldran till andra bärande axlar. Betecknande nog har *axel* 'tvärstång etc.' åtminstone fakultativt fått akut accent trots ursprunglig tvåstavighet, säkerligen under inverkan från *axel* 'skuldra'. Allén (1978) ger fler exempel av samma slag.

I ordböcker måste praktiska hänsyn (konsekvens, lättillgänglighet och klarhet, utrymme m.m.) vägas mot de teoretiska kraven. En etymologisk uppdelningsprincip kan faktiskt i långa stycken ge ett gott s y n k r o n i s k t r e s u l t a t, nämligen i den mån etymologisk härledning sammanfaller med produktiv morfologi. Principiellt är förstås varje historisk indelning psykologiskt otillfreds

ställande, även då det endast gäller att upprätta ordboksenheter. Ett bättre alternativ är enligt vår mening den lemma-lexem-modell som vi tillämpar i projektet Lexikalisk databas. Modellen finns beskriven i Allén (1978), mera som utkast i Ralph, Järborg och Allén (1977). Enligt modellen etableras den lexikaliska enheten först med hänsyn till form och funktion, vilket ger lemmat, därefter med hänsyn till betydelse, vilket ger lexemet. Gemensamma formella och funktionella uppgifter redovisas i lemmadelen, specifika betydelseuppgifter i lexemdelen. Enheten utgörs av en lemmadel j ä m t e en lexemdel.

Lemmat, såsom det definieras av Allén (1967 m.fl. ställen, särsk. i inledn. till NFO 1), utnyttjades ursprungligen som operativ enhet i en kvantitativ korpusundersökning på datamaskinell basis (NFO 1-4). När det nu läggs till grund för en definitionsordbok är det delvis för att det tillsammans med lexemet möjliggör en klar och konsekvent presentationsform. Men vi anser också att ords uttryckssida hör till det som är psykologiskt relevant. Något tillspetsat kan vi påstå att uteslutande form-funktion är minst lika psykologiskt relevant som etymologi. Ordet *balja* kan i svenskan ha betydelserna 'kar' respektive 'fröskida'. I ISO (1977) ger olika historiskt ursprung upphov till två olika enheter. Lemma-lexem-modellen ger en enhet med två skilda betydelser, eftersom båda substantiven *balja* har samma formella och funktionella egenskaper (se äv. Allén 1978). Språkpsykologiskt är det lika rimligt att uppfatta de två enheterna *balja* som varianter av någon grundbetydelse 'behållare' som att hålla isär dem. Den etymologiska information som hänför sig till respektive betydelse är för övrigt mycket specifik (Hellquist 1966); den trotsar all normal intuition och kräver språkhistoriska specialkunskaper. Nu är det kanske inte meningsfullt att betona den språkpsykologiska aspekten alltför starkt så länge det bara handlar om ordböckers presentationssystem. Det är helt klart att även lemman torde kunna associeras till varandra, och lika uppenbart knyter sig vissa formella egenskaper snarast till enskilda lexem. Av en idealisk lexikonmodell, som alltså inte nödvändigtvis behöver svara mot en ordboks krav, bör även sådana relationer som de just antydda framgå.

När lexikonet fokuserats i den nyare teoretiska lingvistiken, har det vanligen varit Chomskys (ursprungligen Bloomfields) lexikonbegrepp man arbetat med: lexikonet uppfattas som en minimal lista över oregelbundenheterna, det unika i språket, dvs. det som inte kan genereras med regler. I den generativa grammatiken sätts de lexikaliska enheterna in i redan genererade strukturer. De lexikaliska enheterna blir med ett sådant synsätt ungefär liktydiga med morfem. Nu är det uppenbart att människan inte lagrar enbart morfem i sitt lexikaliska minne. Särskilt tydligt framgår detta av egennamnen. Skulle vi "komma ihåg" former som {etiop}, {hondur}, {jav} och {ungr} för *Etiopien-etiopier-etiopisk*, *Honduras-honduran-honduransk*, *Java-javanes-javanesisk*, *Ungern-ungrare-ungersk*? Särskilt nationsnamnens former är här helt oförutsägbara. Troligen har människan en inbyggd morfologisk kompetens, men den fungerar säkert på ett helt annat plan.

Man kan märka en tendens till att avfärda egennamn från den lexikaliska diskussionen som triviala och inom språket extrema fall. Då skall man minnas att språkstatistiskt är namnen långt ifrån marginella. I det material på drygt en miljon löpande ord som beskrivits ur olika synvinklar i NFO 1-4 förekommer nästan 11 000 namn på knappt 45 000 ställen. Således är ca 4.5 % av orden i den löpande texten egennamn. Lexikaliskt stiger proportionen till 15 %; totalt har ca 71 000 lemman framanalyserats. Om en så stor del av lexikonet måste lagras som hela ord finns det egentligen ingen anledning att anta något annat än att resten av det lexikaliska materialet lagras i ordform på samma sätt.

I ordböcker lagras onekligen lexikoninventariet normalt i ordform. Ordböcker förlorar emellertid i psykologisk relevans genom de inneboende presentations- och organisationsproblemen: linearireten, den orättmätiga betoningen av alfabetisk ordning, själva den grafiska bildens dominans etc. Vilken uppslagsform man väljer dikteras också till stor del av traditionen. Såväl det lemma som skapas vid lemmatisering som den kanoniska form som väljs som uppslagsform kan representera högst olika entiteter, och förhållandet mellan enhetens realiseringsformer kan variera betydligt. Hur förhåller sig exempelvis en defekt flexionsserie till en fullt utbyggd? Till en uppslagsform *fort* anmäler sig komparationsformerna *fortare-fortast* på ett naturligt sätt, men uppfattas

serien *hellre-helst*, som saknar positivform, som en homogen enhet på samma sätt som den kompletta serien *fort-fortare-fortast*? Kan någondera formen *hellre* eller *helst* anses vara primär framför den andra, såsom kanske *fort* är det jämfört med *fortare-fortast*?

Över huvud taget är paradigmen utfyllda i mycket olika utsträckning för olika ord. En principiell skillnad föreligger mellan sådana paradigm som har stabila luckor och sådana som åtminstone potentiellt kan uppvisa kompletta system. Det finns åtskilliga fall av ofullständiga komparationsserier som visar partiell överensstämmelse med fullständiga, såsom

fort	fortare	fortast	(adverb)
tung	tyngre	tyngst	(adjektiv)
—	hellre	helst	(adverb)
—	sämre	sämst	(adjektiv el. adverb).

Man kan här inte märka någon som helst tendens till nybildning av "analogiska" positivformer som **hell*, **säm*. I NFO 2, som bygger på ett material från 1965, saknar ordet *resurser* på ett liknande sätt singularformer. I detta fall uppfattas emellertid *resurs* spontant som en möjlig singularform. Intuitionen stöds av ett likartat men tio år yngre material, Press 76, där singularformen är belagd:

NFO 2 (material från 1965)		Press 76	
RESURS	121	RESURS	140
s:a sin	*0*	resurs	5
resurser	104	resursen	1
resurserna	17	*s:a sin*	*6*
s:a plu	*121*	resurser	109
		resurserna	25
		s:a plu	*134*

Ordens inherenta betydelse bestämmer naturligtvis i mycket hög grad de olika böjningsformernas inbördes styrkeförhållanden. Så är delvis fallet med *resurs-resurser*, där den plurala eller kollektiva betydelsen ligger nära till hands. På motsatt vis förekommer ord som *värld* och *liv* oftast i singularis, såsom också framgår av siffrorna i NFO 2:

VÄRLD	570	LIV	560
värld	136	liv *sin*	278
världen	310	livet	205
världens	115	livets	39
s:a sin	*561*	livs *sin*	18
världar	9	*s:a sin*	*540*
s:a plu	*9*	liv *plu*	19
		livs *plu*	1
		s:a plu	*20*

Mellan dessa extremer, vars fördelningsförhållanden tycks bero mest på ordens betydelse, återfinns rader av mindre självklara fall. I NFO 2 har *problem* och *väg* totalfrekvenser (559 resp. 552) som är jämförbara med såväl varandra som dem för *värld* och *liv*. Medan *problem* visar relativt jämn fördelning mellan singularis och pluralis (273-286) förekommer *väg* betydligt oftare i singularis än i pluralis (442-103). Även i fråga om andra grammatiska kategorier kan man urskilja individuella profiler för orden. Genitiv är en på det hela taget sparsamt företrädd kategori i svenskan. För enstaka ord bryts dock det vanliga mönstret, som då genitiven är den nästan dominerande realiseringsformen för lemmat *slag*. Man kan jämföra *sak* som har samma frekvens som *slag* men frekvensmässigt negligibla genitivformer.

SAK	453	SLAG	453
sak	183	slag *plu*	69
saken	149	slag *sin*	133
saker	104	slagen	1
sakerna	6	slaget	50
s:a grundform	*442*	*s:a grundform*	*253*
sakens	8	slags	13
sakernas	3	slags	187
s:a genitiv	*11*	*s:a genitiv*	*200*

Fraseologiska konstruktioner kan bidra till att frekvensen höjs för enskilda böjningsformer. Singularformen *väg* understöds av mer eller mindre metaforiska uttryck som *på väg*, *ta vägen*, *vara i vägen* m.m., och genitiven *slags* får sin höga frekvens delvis genom uttryck som *ett slags*, *något slags*. Fraseologi kan slå helt olika inom en klass. Följande adverb, som i NFO 2 har ungefär samma frekvenser (från 297 till 346), ingår enligt NFO 3 i mycket varierande utsträckning i fraser (från 2 till 239).

	frekvens	varav i fraser
verkligen	297	2
däremot	296	8
dessutom	346	13
knappast	307	39
ytterligare	291	99
snart	288	107
alltför	288	144
bort	298	179
ner	286	183
helst	303	215
tillbaka	333	239

Klasser med starkt inskränkt flexion, såsom adverben, kan således vara mycket heterogena med hänsyn till vilken benägenhet de ingående orden visar till att förekomma i fraser. Vi har tidigare sett att enstaka böjningsformer kan framhävas inom paradigmet genom fraseologins inverkan. Härigenom ökas också asymmetrin mellan olika medlemmar av en ordklass. Ovan jämfördes adverben *fort*, inklusive komparationsformerna *fortare-fortast*, och *hellre-helst*. Förutom att den ena serien saknar positivform markeras diskrepansen mellan de båda serierna av den roll fraseologin spelar.

	frekvens	varav i fraser
fort	42	28
fortare	7	3
fortast	2	2
hellre	40	8
helst	303	215

Fort förekommer uteslutande i relativt triviala fraser såsom *för fort*, *lika fort*, *så fort* m.m., *fortare än*, *fortast möjligt*, där *fort* är syntaktiskt utbytbart mot snart sagt vilket adverb som helst. Beträffande *helst* däremot förekommer det hela 209 gånger i den stelnade kombinationen *som helst*, som i uttryck som *vad som helst*, *vem som helst* etc. markerar indefinit betydelse.

Tendens till frasbundenhet är avgjort en viktig lexikalisk uppgift. Det är också tydligt att denna uppgift hänför sig till enskilda ordformer snarare än till den abstrakta enhet som ett ords samtliga böjningsformer bildar tillsammans. I själva verket finns det en del tecken som tyder på att det är ordens böjningsformer man lagrar. Barn är ofta mer direkt medvetna om specifika böjningsformer än om relationerna mellan dem. Det tycks vara en mognadsprocess att vi går från syntagmatiska associationer till paradigmatiska (Salus och Salus 1978). Om människan bygger upp sitt ordförråd genom att först observera enskilda böjningsformer i kontext som tills vidare lagras undan för att först senare associeras till varandra, bör den lexikaliska lagringsformen bära spår av denna process. Frekvenser, som här i viss mån tagits till utgångspunkt för resonemanget men som i språkteoretiska sammanhang ofta nedklassas som ett "performansfenomen", blir därmed ett viktigt begrepp, eftersom frekvenser sedan länge har ansetts vara av grundläggande betydelse för inlärningen. Robinson

nämner exempelvis frekvens som den viktigaste grunden för association (Robinson 1932).

Ordböcker och lexikonmodeller brukar båda vara underkastade stränga ekonomiprinciper. För ordböcker finns det materiella skäl, men lexikonets ekonomi får ses som ett resultat av språkteorins allmänna ekonomitänkande. Kravet på ekonomi och generaliseringar är ofta ett effektivt hjälpmedel i den vetenskapliga analysen, men det är egentligen inget som säger att den mänskliga hjärnan skulle domineras av någon övergripande ekonomiprincip. Tvärtom verkar åtminstone lexikonet vara ganska redundant i sin organisation. Man kan bara tänka på idiomen, som har en perfekt syntaktisk struktur med tolkbara ingående delar men som dessutom fungerar semantiskt på ett fullkomligt arbiträrt sätt. Sådan dubbelhet svär mot varje ekonomiprincip.

Det är troligt att såväl produktionsinriktade som perceptionsinriktade fenomen sätter sina spår i det mänskliga lexikonet. Kanske betonas den semantiska sidan vid perception och upplagring i minnet, medan formen aktualiseras mer vid produktion och framtagning ur minnet. Det mesta av den information man mottar tycks lagras undan rent semantiskt så att den exakta ordalydelsen i meddelandet endast kan rekonstrueras med möda, om alls. Vid produktion, å andra sidan, är det exakta uttrycket viktigt. Det ovan nämnda *slag* kan i de flesta kontexter inte ersättas av det i stort sett likbetydande *art*.

Den enda realistiska modellen av det mänskliga lexikonet torde vara ett flerdimensionellt nätverk som innehåller många sorters information, om betydelse, om form, om funktion osv., med länkar som sammanhåller böjningsformer, synonymer, antonymer, som förbinder fraser med de i dem ingående orden såväl som med synonyma uttryck med annan form osv. Det säger sig självt att en mängd tidskrävande lingvistiska undersökningar behövs innan en sådan modell kan ta form. Lika uppenbart är det att de programmerings- och implementeringstekniska problemen är svårlösta om modellen skall kodas in på datamaskin. I all anspråkslöshet är det ändå principiellt en sådan modell vi har för ögonen när vi bygger upp den lexikaliska databasen i vårt projekt. Informationsmängden kan utökas successivt och måste kanske delvis omstruktureras,

länkningen får sofistikeras efterhand. Det viktiga som skall slås fast här är att en psykologiskt rimlig lexikonmodell i mycket överensstämmer med de databasstrukturer i form av länkade nätverk som nu utvecklas på flera håll. Därigenom understryks ytterligare vikten av att datalingvister engagerar sig i det lexikologiska arbetet.

Referenser

Allén, S. 1967. Förhållandet mellan tal och skrift. I: Allén et al., Språk, språkvård och kommunikation. Lund 1967.

Allén, S. 1978. Lexical entry, linguistic sign, and lexical data base. Föredrag vid 7th International Conference on Computational linguistics. Bergen 1978. Utkommer i förhandlingar från konferensen.

Hellquist, E. Svensk etymologisk ordbok. 3 uppl. 4y tr. Lund 1966.

ISO = Illustrerad svensk ordbok. 3 uppl., 3 tr. Stockholm 1977.

Katz, J.J. och J.A. Fodor. 1963. The structure of a semantic theory. I: Language 39. Omtryckt i: The structure of language, utg. av J.A. Fodor och J.J. Katz. Englewood Cliffs 1964.

McCawley, J.D. 1968. The role of semantics in a grammar. I: Universals in linguistic theory, utg. av E. Bach och R.T. Harms. New York etc. 1968.

NFO = Nusvensk frekvensordbok. 1-3 (1970-75), 4 (utkommer). Stockholm.

Press 76. Konkordans över ett tidningsmaterial från 1976. Språkdata, Göteborg. [Radskrivarutskrift.]

Ralph, B, J. Järborg och S. Allén. 1977. Svensk ordbok och Lexikalisk databas. Förstudierapport. Språkdata, Göteborg. [Stenc.]

Robinson, E.S. Association theory today. New York 1932.

Salus, M.W. och P.H. Salus. 1978. The acquisition of opposites and the structure of the universe. Papers in language use and language function 1. Scarborough College, University of Toronto. [Stenc.]

Denna artikel har delvis kunnat utarbetas inom ramen för projektet Lexikalisk databas.