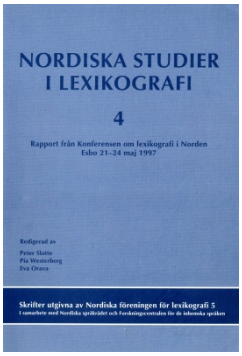
NORDISKE STUDIER I LEKSIKOGRAFI

Titel: Återanvändning av ordboksmaterial - mål och metoder

Forfatter: Inger Hesslin Rider

Kilde: Nordiska Studier i Lexikografi 4, 1997, s. 181-187
Rapport från Konferens om lexikografi i Norden, Esbo 21.-24. maj 1997

URL: http://ojs.statsbiblioteket.dk/index.php/nsil/issue/archive



Søgbarhed
Artiklerne i de ældre Nordiske studier i leksikografi (1-5) 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.

Inger Hesslin Rider

Återanvändning av ordboksmaterial – mål och metoder

The advent of new technology has made it possible to make more efficient use of existing dictionary material in the elaboration of new works. The paper describes three authentic instances: using the source-language data base of a bilingual dictionary to create a dictionary with a new target language, retrieving and updating text not available in digital form, and reversing a bilingual dictionary. It describes the reasoning behind each case and discusses advantages and problems as well as the demands these methods put on the lexicographer.

Återanvändning av ordboksmaterial är inget nytt fenomen, åtminstone om man därmed avser återanvändning av andras material. Det har alltid funnits en viss rundgång i ordboksbranschen. Man har 'låtit sig inspireras' av olika redan existerande material i större eller mindre omfattning. Det kan förvisso vara förståeligt, med tanke på den tid och det arbete en ordbok kräver, men det finns en gräns som inte får överskridas. En gräns där det blir olagligt eller i alla fall omoraliskt.

Det kunde vara ett ämne för en egen betraktelse. Men den här artikeln ska handla om återanvändning av ett förlags eget material. Jag vill resonera kring varför och hur ett förlag väljer att återanvända ordboksmaterial, beskriva några olika tillvägagångssätt och belysa med exempel från några konkreta projekt. Jag kommer därvid också att beröra de krav som arbetssättet ställer på redaktörerna.

I kallelsen till årets NFL-konferens stod apropå återanvändning frågan "Tjänar tillvägagångssättet användarna eller snarare lexikografen och förläggaren?". Det är en fråga som lockar till diskussion. Jag ska försöka illustrera min övertygelse att förlagen och de lexikografer som arbetar där inte står i något motsatsförhållande till användarna utan tvärtom. Det affärsdrivande förlaget måste hålla sig på användarens sida. Att skapa material för användarna – material som kommer ut, används och hela tiden förbättras – är det enda val ett förlag har om det vill fortsätta att finnas.

Datorisering och annan ny teknik har gjort återanvändning möjlig. Skälen till att man återanvänder ordboksmaterial är naturligtvis en viss tids- och kostnadsbesparing, men den är inte det primära och inte heller så stor som man kanske kunde tro. Metoden ger också möjlighet att bevara ordboksmaterial som annars skulle gå förlorade. Till sist ger den utrymme för kvalitetshöjning genom att redaktören kan fokusera på nya saker i texten. Jag kommer här att redovisa tre metoder: att utifrån källspråksbasen i en tvåspråkig ordbok skapa en ordbok med annat målspråk, att använda optisk läsning och parsing för att göra ett existerande material datamässigt hanterbart och till sist att vända en tvåspråkig ordbok.

Förutsättningar

För att ett förlag ska kunna återanvända ett material krävs tillgång till ett datasystem som är flexibelt och som medger att redaktörerna gör experiment och modifieringar utan extern hjälp.

Däremot behöver man inte kunna göra *allt* på redaktionen – bl.a. av praktiska skäl eftersom vissa operationer kräver mycket stor datorkapacitet.

Det datasystem vår redaktion använder heter Compulexis. Vi började använda systemet i dess ursprungliga version i mitten av åttiotalet. Sedan dess har det utvecklats och blivit allt mer kraftfullt och användbart. Compulexissystemet är välkänt för många, men en kort beskrivning kan ändå vara på sin plats. Som redigeringssystem är det starkt strukturerat, något som underlättar användningen och har avgörande betydelse för den slutliga överföringen till sättfiler eller omvandling för utgivning i elektronisk form. Strukturen byggs om för varje given ordbok, eller efter förlagsspecifika önskemål. All text som förs in märks med koder, 'tags' (se fig.1). Koden talar om vilken typ av innehåll som finns i det följande fältet. Den talar också om om fältet tillhör käll- eller målspråk. Systemet tar hand om mycket av det som förut krävde mycket tid och manuell kontroll, t.ex. alfabetisering enligt ett givet språks principer, den inbördes ordningen mellan momentsymbolerna (siffror och bokstäver), skiljetecken mellan kategorier, vilken stilsort som ska användas för en viss informationstyp, vilka förkortningar som är tillåtna m.m. Utöver koderna behövs inga stilmarkeringar eller specialtecken. Redaktören gör själv sökningar, modifieringar och beräkningar kategorivis. I Compulexissystemet kan man också i varje ögonblick se sin kodade text som typograferad text med alla tecken på plats (se fig. 2). Det är alltså ett helt utvecklat WYSIWYG-system – man kan hela tiden se vad det är man skapar.

För allt det som ligger efter det att materialet är klart på redaktionen finns en väl uppdragen bana till sättning och framställning av tryckoriginal. Dessutom kan materialet förvandlas till enklare textfiler eller andra system.

Fallstudie 1: återanvändning av källspråkssidan i en tvåspråkig ordbok

Det första fall jag vill beskriva är ett exempel på hur man kan återanvända källspråkssidan ur en tvåspråkig ordbok. Vi hade i sortimentet en svensk-spansk ordbok som var så pass föråldrad att det var olämpligt att försöka bygga vidare på den. Den var för liten, stickordsförrådet var inte tillräckligt aktuellt och översättningarna hade ibland en lätt ålderdomlig klang. Däremot hade vi en nyskriven svensk-fransk ordbok. Det här skulle bli ett av de första projekt där vi verkligen drog nytta av den nya teknikens fördelar. Vi bedömde det som en stor fördel att frigöra oss från den gamla ordbokstexten. Alla som sysslat med praktiskt ordboksförfattande känner säkert till problematiken: på något sätt bär det ofta emot att stryka och man tenderar att lägga till istället! Eftersom allt i Compulexissystemet är märkt så att man kan avgöra vilken kategori det tillhör och om det hör till käll- eller målspråket, är det lätt att identifiera det man vill använda. Man kan t.ex. välja att behålla uppslagsord, ordklassbeteckning, ämnesområdesbeteckningar, semantiska upplysningar och fraser, eventuellt också strukturen med momentsymboler. Det man får fram på detta sätt är ett skelett, ett utkast – men det är inte hälften av en ordbok och det ska man ha väldigt tydligt klart för sig. Varje artikel, varje ord måste skärskådas av kunniga redaktörer med stor medvetenhet om skillnaderna mellan det gamla och det nya målspråket.

Om man jämför de färdiga ordböckerna (svensk-fransk och svensk-spansk) är skillnaden uppenbar redan på stickordsnivå – där den svensk-franska ordboken bara har fyra uppslagsord, ger den svensk-spanska 16 sammansättningar och avledningar:

sv-fr	*sv-sp*
Spanien	**Spanien**
	spanienkännare
spanjor	**spanjor**
spanjorska	**spanjorska**
spansk	**spansk**
spanska	**spanska**
	Spanskamerika
	spanskamerikan
	spanskamerikansk
	Spanska ridskolan
	spanskfientlig
	spanskfödd
	spanskspråkig
	spansk-svensk
	spansktalande
	spanskvänlig

För de partier som på motsvarande sätt anknyter till *fransk* är förhållandet det omvända. När det gäller en artikel som helhet varierar behovet av omarbetning. Innehållsordens artiklar kan ofta byggas på den gamla källspråksstrukturen, medan funktionsorden i princip måste formuleras om helt. I den här typen av arbete blir det ofta uppenbart hur målspråket påverkar källspråket. I nedanstående artikelpar är den svensk-franska artikelns infinita konstruktioner inte funktionella för motsvarande svensk-spanska uttryck. Jämför:

> **fotsvett** ***s, ha*** ~ transpirer des pieds; ***lukta*** ~ om person sentir des pieds (*Norstedts svensk-franska ordbok*)
> **fotsvett** ***s, han har*** ~ le sudan los pies; ***han luktar*** ~le huelen los pies; ***det luktar*** ~ huele a pies (*Norstedts svensk-spanska ordbok*)

Vad är då fördelen med att utgå från en befintlig ordbok? Tidsvinsten är som sagt var inte huvudskälet. När vi startade projektet hade vi nog förväntningar på att metoden skulle halvera tidsåtgången, men det visade sig vara helt fel. En tidsbesparing på 10–15% kan man kanske räkna med, men den inbesparingen ligger i huvudsak på inkodningen. Fördelen är dels att man inte behöver skriva om sådant som redan finns, som fungerar i den nya kontexten och som är rättstavat – varje tangenttryckning är ju annars en potentiell felskrivning. Och i stället för att uppfinna det berömda hjulet och än en gång samla ihop ett svenskt ordförråd kan redaktörerna ägna sin tid och kraft åt att skärskåda och finslipa det, och åt det viktigaste, nämligen att finna adekvata översättningar.

Fallstudie 2: bevara och vidareutveckla gammal sättning

Det andra fall jag vill ta upp uppstod i en problemsituation. En ordbok var utgången på förlaget men ständigt efterfrågad, väntelistorna på antikvariaten var långa. Boken, Gullberg: *Svensk-engelsk fackordbok*, utkom i sin andra upplaga 1977. Det var fortfarande i blysättningens dagar.

Ordboken trycktes sedan om ett antal gånger, med allt sämre tryckteknisk kvalitet, och anmäldes till sist som utgången. I takt med att de existerande exemplaren började falla sönder ökade pressen på förlaget att ge ut den igen. Det rör sig om ett stort material på ca. 170.000 termer, fördelade på huvudbok och supplement, tillsammans ca. 1.700 sidor. Författaren hade dessutom samlat på sig ett stort ytterligare material, skrivet på ca. 200.000 manuskort. Av allt detta skulle vi göra en modern ordbok. Att fotografera av boken och lägga till ännu ett supplement var inget acceptabelt alternativ. Läsbarheten skulle bli mycket dålig och att placera tillägg i supplement är ett föga användarvänligt förfaringssätt.

Den väg vi valde blev denna: Ett exemplar av boken skickades till optisk läsning. OCR-läsning (*Optical Character Recognition*) har numera följts av *ICR-läsning*, där I:et står för *Intelligent*. De gamla OCR-systemen kunde bara arbeta med en viss font i taget. ICR-systemet kan tränas för att klara olika tecken, inklusive sådana som åstadkoms av trasiga blytyper, och är mycket effektivt vad gäller att klara många olika grader och stilsorter samtidigt. Det kan dessutom identifiera långa typiska strängar och 'tvätta' återkommande feltyper.

Efter den optiska läsningen av den gamla boken och supplementet kördes bokens text därefter ut som textfiler med fontskiftsmarkeringar. Filerna lämnades till Compulexis för parsing, märkning med Compulexiskoder. Eftersom ordbokens artikelstruktur var relativt enkel kunde parsingen göras med god träffsäkerhet.

Medan detta arbete med ursprungsmaterialet pågick, hade en projektgrupp gjort ett urval på ca 30.000 av de 200.000 nya manuskorten. Av dem var ca 20.000 helt nya ord, dvs. ord som inte fanns i ordbokens huvudtext eller i supplementet. Dessa skrevs in i en egen ordbas i Compulexis, medan kort som innebar tillägg eller ändringar till befintligt material lades åt sidan för senare inredigering.

När parsingen var klar sorterades huvudbok och supplement ihop alfabetiskt med den mindre ordbasen. Varje artikel försågs med en s.k. 'flagga', en kod som talade om huruvida artikeln kom från huvudboken, från supplementet eller från tilläggsorden. När ett uppslagsord förekom i två eller alla tre av dessa kategorier placerades artiklarna intill varandra i för senare redigering. Eftersom ordboken varit redigerad med många fall av s.k. 'nästen' som innehöll krok (tilde) eller bindestreck, hade sammansättningar och avledningar dessförinnan expanderats till sin fulla form för att kunna sorteras. För säkerhets skull hade varje uppslagsord också fått en 'adresslapp', ett fält som visade av vilka orddelar det hade konstruerats.

Och här befinner vi oss nu: vi har materialet digitalt lagrat, och håller på med det redaktionella arbetet. Vi kan söka, redigera, beakta de 10.000 ändringskorten och göra konsekvenskontroller. Den tredje upplagan kommer om några år. Det blir ett verk med modern lättläst typografi, utan supplement; en utökad och lätt reviderad version av föregående upplaga. Det stora, viktiga steget är att denna många fackköversättares favorit nu kan leva vidare och förädlas. Det hade inte gått utan 'återanvändning'.

Fallstudie tre: vändning

Det tredje fallet jag vill ta upp är vändning av en tvåspråkig ordbok. Redaktionen har, med Compulexis hjälp, experimenterat och förbättrat sina kunskaper om vändning i sex–sju års tid. Vi har funnit att det centrala är att hitta rätt ambitionsnivå, en balans mellan att få så mycket 'färdigt' material som möjligt och att kunna eliminera sådant som blir alltför snårigt och oanvändbart.

En enkel vändning kan illustreras med följande ordboksartikel:

apelsin *s* Apfelsine -n *f*, Orange -n *f*

Artikeln består av följande fält: uppslagsord, ordklass, översättning, böjning och genusbeteckning, översättning, böjning och genusbeteckning. Datorn börjar överst i artikeln och letar sig fram till det första fält som är kodat som översättning; på vägen lagras uppslagsordet, eventuell ordklassbeteckning och eventuell ämnesbeteckning. Nu skapas en ny artikel där översättningen blir uppslagsord. Därefter kontrollerar datorn om översättningen följdes av en genusbeteckning. Om den inte gjorde det kan man anta att ordklassbeteckningen ska vara densamma för båda språken. Finns genusbeteckning vid översättningen definieras den om till ordklassbeteckning och slår ut beteckningen *s* (substantiv). Det ursprungliga uppslagsordet blir översättning.

Apfelsine -n *f* apelsin

Därefter börjar processen om; eftersom ursprungsartikeln innehöll flera översättningar återgår man till den och **Orange** blir ett nytt uppslagsord, annars går datorn vidare till nästa artikel.

I artiklar som innehåller långa översättningar, t ex frasöversättningar, söker datorn i första hand efter ettordsöversättningar som den stött på tidigare i artikeln för att låta frasen följa med till den nya artikel som skapas. Innan datorn kan hantera mer komplicerade artiklar måste ett förarbete vara gjort – tilde och bindestreck i fraser måste ha byggts ut till relevant form m.m. Vidare måste det finnas en lista över ord som är 'grus', d.v.s. innehållsligt mer eller mindre tomma, ord som fraser i allmänhet inte bör föras till.

När hela ordboksmaterialet blivit behandlat på detta sätt vidtar en ny datoromgång där det nya, vända materialet samlas i alfabetisk ordning och struktureras så att alla artiklar som nu har fått samma uppslagsord – dvs f.d. översättningar – förs samman till en artikel. Det är viktigt om det ska bli hanterbart och överblickbart. Hur materialet ställs upp följer principer knutna till ordklassbeteckning, ämnesområde m.m., som det här skulle föra för långt att gå in på.

En brist i ett vänt material är just att det är vänt – dvs att det speglar det gamla källspråkets värld. För att försäkra sig om att det nya materialet blir balanserat och innehåller de kulturspecifika orden från det *nya* källspråket kan det vara klokt att motköra mot en lämplig ordbas.

Det säger sig själv att en text som skapats på detta sätt kräver både en noggrann och kritisk granskning, och ett omfattande efter- och kompletteringsarbete. Det faktum att man inte ska förvänta sig ett färdigt material innebär på intet sätt att metoden är oanvändbar. Generellt kan sägas att den idealiska ordboken för vändning är en stickordstät ordbok med enkel struktur – om det finns en eller många översättningar i en artikel spelar däremot ingen roll. Ju mer 'teknisk' eller fackbetonad en ordbok är, desto bättre blir resultatet av vändningen. En fackordlista kan mycket väl vändas med bra resultat till nästan 100 procent medan andelen potentiellt 'klara' artiklar i en vanlig ordbok är mycket mindre. Dessutom kan man säga att ju större ordboken är, desto större blir andelen artiklar som inte kräver mycket ytterligare behandling. Det beror naturligtvis på att basordförrådet i ett språk är det mest komplicerade och svårast att vända. Basordförrådet utgör en stor andel av uppslagsorden i en liten ordbok, medan det tvärtom är en liten del av en stor ordbok. Våra vändningsrutiner innehåller en logik för att undanta denna typ av artiklar och skapa dem helt manuellt.

Den vanligaste invändningen mot att vända ordboksmaterial är att även om X på källspråket översätts med Y på målspråket, är det inte alldeles säkert att det blir korrekt i omvänd riktning.

Det stämmer, men faktiskt inte generellt utan bara i ett ganska litet antal fall. I verkligheten finner man att de allra flesta paren ord/översättning fungerar i båda riktningarna. Den stora vinsten med vändning är att det är en metod för att hitta material som kanske annars inte skulle hittas – i bästa fall dessutom korrekturläst och klart. Det ger också en stor möjlighet att få balans mellan ordbokspar. En nyttig spinoff-effekt är att vändningen gör oss uppmärksamma på sådant som inte är helt lyckat i det gamla materialet.

Oavsett vilken arbetsmetod man använder – vare sig man skriver helt nytt, gör nya upplagor eller använder modern teknik i en eller annan form – är A och O att allt material som åstadkoms alltid gås igenom av redaktörer, redaktörer som kan sin grammatik, sina lexikografiska principer och som kan redigera och som vet för vilken målgrupp materialet görs. Redaktören förblir hjärnan oavsett hur många elektronhjärnor han eller hon har till sin hjälp.

Avslutning

Jag har gett några exempel på vad den nya tekniken gör tillgängligt för oss: vi kan bygga upp ett modernt ordförråd och hålla det levande som ett centralt och separat projekt, snarare än att återuppfinna det för varje enskild ordbok, vi kan använda optisk läsning och parsing för att 'rädda' stora material som annars skulle gå förlorade och vi kan vända material för att få nya perspektiv. Vi kan använda dessa metoder i kombination och det finns många fler metoder som ger andra fördelar. Gemensamt för dem alla är att de bara är en liten bit på vägen och att de ställer stora krav på de redaktörer som arbetar med dem. Det viktiga är inte att de ger tidsvinster utan att de ger möjlighet att fokusera på det viktiga, nämligen innehållet i ordböckerna, så att framtidens ordböcker kan bli de levande och rörliga dokument som hör IT-åldern till.

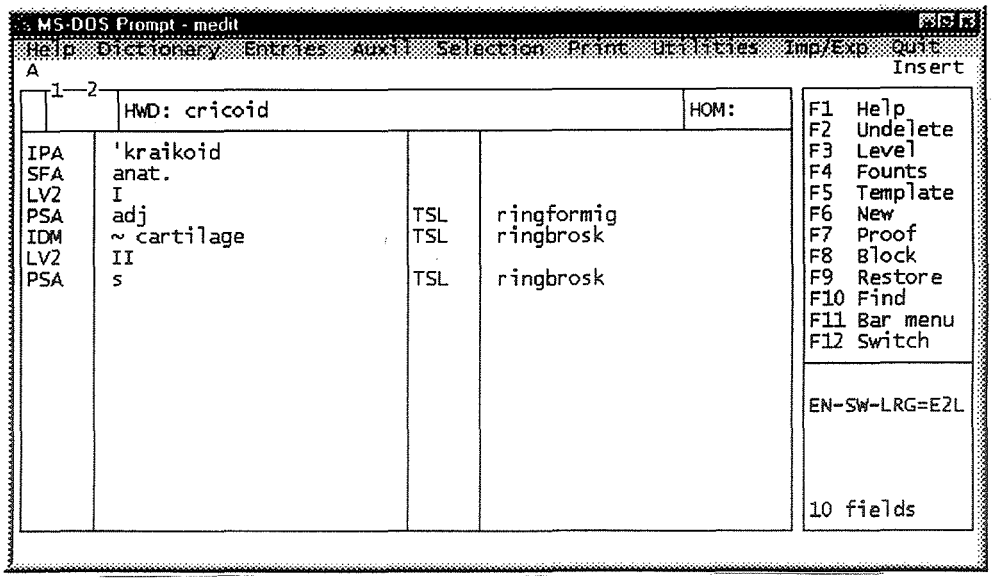


Fig. 1 *En artikel i Compulexissystemets "Forms Mode", med koder som anger vad nästa fält innehåller.*

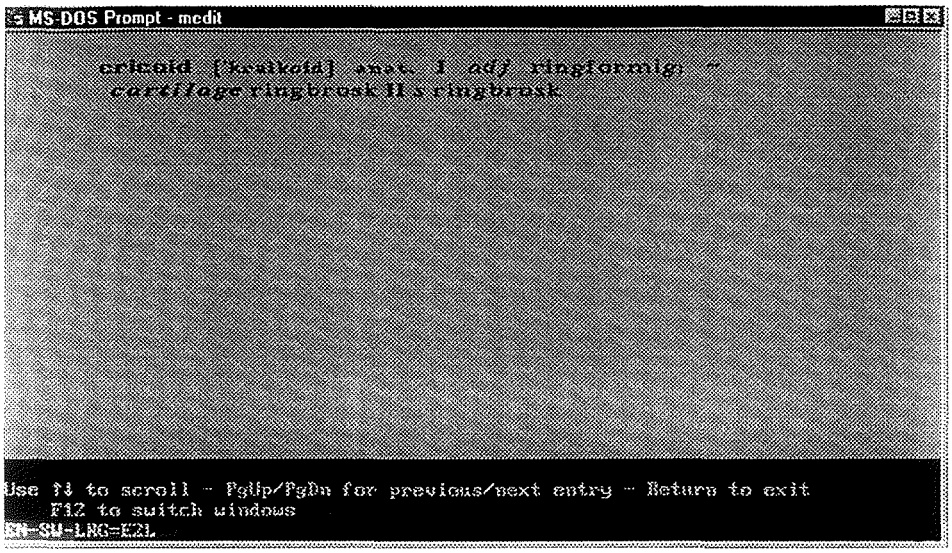


Fig. 2 *Samma artikel i "Proof Mode", d.v.s. som typograferad text.*