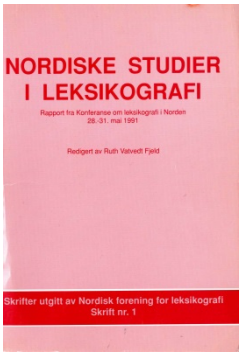
NORDISKE STUDIER I LEKSIKOGRAFI

Titel: Lexikografiska principer för alfabetisk filering med dator

Forfatter: Rolf Gavare

Kilde: Nordiske Studier i Leksikografi 1, 1992, s. 184-189
Rapport fra Konferanse om leksikografi i Norden, 28.-31. mai 1991

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



Betingelser for brug af denne artikel
Denne artikel er omfattet af ophavsretsloven, og der må citeres fra den. Følgende betingelser skal dog være opfyldt:

- Citatet skal være i overensstemmelse med „god skik“
- Der må kun citeres „i det omfang, som betinges af formålet“
- Ophavsmanden til teksten skal krediteres, og kilden skal angives, jf. ovenstående bibliografiske oplysninger.

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.

Rolf Gavare

Lexikografiska principer för alfabetisk filering med dator

Den alfabetiska ordningen har sedan århundraden varit vår viktigaste nyckel för att komma åt information via lexikaliska register. När vi lämnar det manuella inordnandet och övergår till automatiska datarutiner, upptäcker vi snart att den traditionella alfabetiseringens principer är långt ifrån så triviala och entydiga att de enkelt kan formaliseras. Detta bidrag försöker ge en liten inblick i alfabetiseringens huvudproblem och beskriver en lexikografiskt baserad modell som kan implementeras i våra datorsystem.

Vad är lexikografisk alfabetisering?

Frågan kan naturligtvis enkelt besvaras med att begreppet avser *ordnandet av ett antal stickord eller flerordiga uppslagsformer i enlighet med alfabetets bokstavsordning*. Detta låter ju mycket enkelt och klart, men tyvärr visar det sig vid närmare eftertanke finnas en mängd delproblem dolda i denna fråga. Låt mig därför ta det hela från grunden och göra en snabb översikt över några av de viktigaste problemen vid alfabetiskt ordnande.

Vad är det egentliga syftet med en alfabetisk ordning?

Huvudsyftet är förstås att läsaren/användaren så snabbt som möjligt skall kunna finna den information som han eller hon söker efter. Sedan flera hundra år har man utnyttjat den inlärda, fasta ordningsföljd som ges av bokstäverna i alfabetet. Den hävdvunna, alfabetiska följden varierar emellertid en hel del från språk till språk och har också varierat genom tiderna — vi kan exempelvis påminna oss att runraderna, futharkerna, ger en helt annan ordning än dagens alfabet. Bokstavsramsan har en mnemoteknisk funktion, men den återspeglar också ofta en del av det aktuella språkets (morfo)fonematiska särdrag. Vad vi förknippar med en korrekt bokstavsordning är dock, nota bene, inte alltid en linjär följd av bokstäver. (Jfr tabell 1.)

Det bör betonas att en alfabetisk ordningsföljds främsta syfte är att oinitierade användare — utan några särskilda anvisningar — snabbt skall kunna finna den information de söker, dvs. den skall finnas på en intuitivt naturlig plats. Det är just detta syfte som bör vara vägledande för lexikografer och andra som upprättar alfabetiska förteckningar.

Var finns problemen?

Så länge det gäller att alfabetiskt ordna enskilda ord med enbart små bokstäver, utan accenter

bokstavsordning

Ordningsföljden i de svenska, isländska, tyska och spanska alfabeten.

Svenska	*Isländska*	*Tyska*	*Spanska*
a, à	a	a, ä	a, á
b	á	b	b
c	b	c	c
d	(c)	d	ch
e, é	d	e	d
f	ð	f	e, é
g	e	g	f
h	é	h	g
i	f	i	h
j	g	j	i, í
k	h	k	j
l	i	l	k
m	í	m	l
n	j	n	ll
o	k	o, ö	m
p	l	p	n
q	m	q	ñ
r	n	r	o, ó
s	o	s (ß=ss)	p
t	ó	t	q
u	p	u, ü	r
v, w	(q)	v	s
x	r	w	t
y, ü	s	x	u, ú, ü
z	t	y	v
å	u	z	w
ä	ú		x
ö	v		y
	(w)		z
	x		
	y		
	ý		
	z		
	þ		
	æ		
	ö		

Tabell 1. Ur Nationalencyklopedin, artikeln *bokstavsordning*.

eller specialtecken av något slag, så är det inga större problem. Komplikationerna uppkommer framförallt vid *alfabetisering av större ordböcker och uppslagsverk, bibliografiska kataloger, personnamnsregister, tekniska termlistor och konkordanser och ordindex till stora, autentiska textmaterial.* Hur skall man t.ex. behandla *skillnaden mellan gemena bokstäver och versaler, diakritiska tecken* (som i *ç, é, ł* och *ñ*), *ligaturer* (som *æ* och *œ*), *digrafer* (som spanskans *ch* eller holländskans *ij*), *bokstavsvarianter* (som *v-w* och *y-ü* i svenskan), *icke-latinska skrivtecken* (t.ex. grekiska bokstäver i vetenskaplig text), *logogram* (*&*, *%*, *§*, etc.), *förkortningar, siffertal (arabiska och romerska), skiljetecken, symboler* osv.? Var kan man

exempelvis förvänta sig att en uppslagsboks läsare söker artiklarna *1,3,5-triazin*, *Henri de Toulouse-Lautrec*, *Le Havre*, *'s-Gravenhage*, *Þingvellir*, *Lübeck*, *α-strålning*, *Karl X Gustav*, *SJ*, *&*, *o.s.v.*, *Cæsar*, *Ærø* och *1984*? Hur skall de nämnda typerna prioriteras sinsemellan och hur skall varje grupp vara rangordnad internt; vilken inbördes ordning är t.ex. mest logisk att ha mellan de olika diakritiska tecknen?

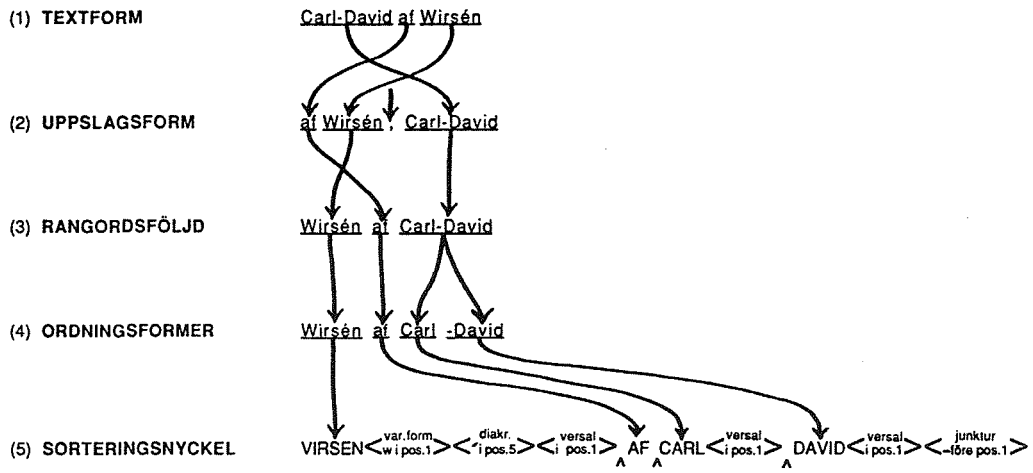
Ett för användaren ofta mycket påtagligt problem gäller också om alfabetiseringen av flerordiga uppslagsformer skall ske *ord-för-ord* eller *tecken-för-tecken* (så att ordgränserna negligeras); det är klart förvirrande när olika principer tillämpas i likartade sammanhang. Skillnaden är förstås särskilt märkbar då uppslagsformens inledande ord är kort. Vid ordvis alfabetisering är det därför t.ex. mindre lämpligt att skriva akronymer med ordmellanrum mellan bokstäverna. Principiellt är en alfabetisering ord-för-ord att föredra om bruket av sär- resp. sammanskrivning är relativt stabilt, som t.ex. i de nordiska språken. Om det däremot vacklar eller varierar fritt kan tecken-för-tecken-principen vara lämpligare. (Det kan dock noteras att t.o.m. Encyclopedia Britannica nyss övergått från sortering tecken-för-tecken till ord-för-ord.)

Krav på algoritmer för datoriserad alfabetisering

Våra grundläggande krav på principer för att åstadkomma en alfabetisk ordningsföljd är att *ordningen skall bli förutsägbar och entydig*. Detta är särskilt betydelsefullt vid datorbaserad sortering. Kravet innebär att vi måste kunna vara säkra på att få exakt samma ordningsföljd om samma material alfabetiseras på nytt och att det inte blir några komplikationer om vi samsorterar alfabetiserade material från olika källor. Omfattande dataregister för informationssökning och urval ställer också stora krav på explicita och entydiga fileringsmetoder och användaranpassade presentationsformer. *All information som finns i uppslagsordens grafiska form måste därför kunna utnyttjas*, om vi skall kunna tillfredsställa höga lexikografiska krav på stringens och konsekvens. Detta har tyvärr de hittillsvarande standarderna inte klarat av. Jag skall här i korthet beskriva hur en — i tillämpliga delar — datorimplementerbar modell för alfabetisering kan vara uppbyggd.

Föreslagen alfabetiseringsmodell

Med utgångspunkt i nedanstående illustration (figur 1) skall jag försöka belysa modellens huvudmoment. Vi utgår från de enheter som vi vill använda som sökbara referenser och som skall ordnas i en intuitivt korrekt, alfabetisk följd. De är i sin ursprungliga form ord eller ordförbindelser, så som de faktiskt återges i löpande text (med normal ordföljd). Denna s.k. *textform* (betecknad som nivå 1 i figur 1) är dock inte alltid den vi är vana att använda som *uppslagsform* (2). En viktig grupp där uppslagsformen traditionellt skiljer sig från textformen är personnamnen. I Sverige och flertalet europeiska länder skrivs i sökregister nutida personnamn med *inverterad följd*, dvs. med efternamn, kommatecken och därefter förnamn. Medeltida och äldre personnamn anges dock i *rak följd*, i likhet med bruket på Island (vad beträffar islänningars namn, ofta till skillnad från utlänningars). I vissa fall förekommer det också att andra, flerordiga namn får inverterad ordföljd (som t.ex. typen *Skandinavien, Brand- och olycksfallsförsäkringsaktiebolaget*).



Figur 1

En invertering av namnformen är dock inte tillräcklig i alla fall, exempelvis när det gäller hanteringen av namnprefix (*af, von* etc.). I bibliografiska förteckningar behandlas namnprefixen enligt mycket komplicerade regler. Exempelvis brukar prefixet *de* negligeras i namn på danskar, norrmän, holländare, flamländare, fransmän, medeltida italienare, spanjorer, portugiser, m.fl., medan det blir sorteringsgrundande för exempelvis svenskar, engelsmän och nutida italienare. Extra problem uppträder naturligtvis för personer som bytt nationalitet. Så här komplicerade regler är absolut inte användarvänliga — de måste klart förenklas, kanske t.o.m. så mycket att samtliga personnamnsprefix endast blir sekundärt sorteringsgrundande.

Bibliotekens katalogiseringsregler ger oss också ett annat exempel på hur delar av uppslagsformen (t.ex. en boktitel) särbehandlas: inledande prepositioner och artiklar och bestämningar av typen *Svenska, Statens, Allmänna, Aktiebolaget* etc. brukar negligeras; eventuellt får de bilda sista sorteringsgrund. (Effekten blir då i princip likvärdig med en invertering av uppslagsformen.)

För att nämna ytterligare en anmärkningsvärd bibliografisk konvention kan vi tänka på att tecknet *&* oftast insorteras som om det vore utskrivet som *och, and, und, et, og* etc., alltefter titelns språk.

Det är således befogat att urskilja vad som brukar kallas en *rangordsföljd* (3), dvs. de element (ord) som den alfabetiska sorteringen skall ta hänsyn till, i en preciserad, inbördes ordning; alltså inte nödvändigtvis varje ord från det första till det sista.

För att undvika sorteringsfel som beror på den mänskliga faktorn bör man vara mycket restriktiv med manuella ingrepp i underlaget för sorteringsnycklarna. Hela sorteringsförloppet efter etablerandet av rangordsföljden kan ske helt automatiskt, algoritmiskt — inga manuella ändringar bör därför tillåtas sedan rangordsföljden fastställts; *all information som sorteringen grundar sig på bör vara en entydig återspegling av uppslagsordens grafematiska form.*

Rangordsföljden är emellertid inte omedelbart användbar som underlag för genereringen av sorteringsnyckeln. Vi måste precisera de explicita teckenföljder som skall

jämföras vid sorteringen. Detta sker i den s.k. *ordningsformen* (4). Här omskrives exempelvis ligaturen *æ* till *oe*, *ß* blir *ss* och *ch* blir i spanska material jämställt med ett enda tecken (en kod), jfr tabellen.

I somliga tillämpningar omkodar man på detta stadium arabiska siffertal till en rent alfabetisk form och romerska ordningstal till ett numeriskt värde (för att få korrekt ordning på bl.a. regentföljder).

För vissa applikationer ersätts nu de ortografiskt korrekta ordningsformerna med fonetiska former eller kanoniska normalformer, som sedan får bilda underlag för sorteringen. (Jfr telefonkatalogens normaliserade efternamnsortering, som samsorterar *Carlson, Carlsson, Karlson, Karlsson* etc.)

Vid *finalalfabetisk sortering*, slutligen, vändes ordningsformernas teckenföljd baklänges.

Sorteringsnyckeln (5) skapas därefter enligt den s.k. *stavningsprincipen*, dvs. sorteringsnyckeln bygger strikt på (de eventuellt modifierade) ordningsformernas stavning.

I korthet kan följande principer fastställas för hur sorteringen bör fungera och därmed för hur sorteringsnycklarna bör vara konstruerade:

> Sorteringen kan ske *ord-för-ord* eller *tecken-för-tecken* (i vilket fall ordmellanrummen negligeras). Trots att skillnaden mellan dessa principer, som nämnts, ofta är mycket påtaglig, finns ingen universellt accepterad konvention eller standard i detta avseende. Svenska telefonkataloger och bibliotekskataloger tillämpar exempelvis ord-för-ord-principen medan uppslagsverk traditionellt tillämpat tecken-för-tecken-principen. (Enligt den förra sorteras således *allmän väg* före *allmänna* och *Svensk uppslagsbok* före *Svenska Akademien*, enligt den senare metoden tvärtom.) För (svenskt) lexikografiskt bruk förordas att sorteringen sker ord-för-ord.
>
> *Vid rent alfabetisk sortering negligeras till att börja med alla icke-alfabetiska tecken; vid ordlikhet får övriga tecken* (siffror och skiljetecken) *fälla utslaget.* (Exempelvis hamnar då *1,3,5-triazin* direkt efter *triazin.*) Vid *alfanumerisk sortering* behandlas siffrorna däremot som jämbördiga med bokstäverna i sorteringshänseende; *1,3,5-triazin* hamnar då bland sifferorden.
>
> Jämförelsen av de alfabetiska tecknen görs i varje ord på fyra nivåer:
>
> 1. Först jämföres orden i *normaliserad* form, där enbart motsvarande alfabetiska grundtecken och logogram beaktas; vi får då exempelvis följden *& a Aabel* ....
> 2. Vid (ord)likhet tas därefter hänsyn till ev. *varianttecken* i orden: *Lybeck Lübeck ... tvist twist tvista twista tvistemål ...* (enl. svensk konvention).
> 3. För det tredje beaktas eventuella *diakritiska tecken*: *cote coté côte côté* ....
> 4. Slutligen ges bokstäver som utmärks med *versal* lägre prioritet än motsvarande gemen: *ma mA Ma MA* ....
>
> Jämförelsen bör alltså normalt ske *ord-för-ord* samt från vänster till höger, så att ord som börjar på ett likartat sätt hamnar intill varandra. (Vi får då exempelvis ordningen *in in- -in -in-* ...; jfr även 3 och 4 ovan.)

Den modell för alfabetisk filering som jag här kortfattat presenterat finns utförligare beskriven

i Gavare 1988. Häri finns också utförliga anvisningar för modellens implementering i datorprogram. (Modellen är underlag för Teknisk norm (nr 34) för svenska statsförvaltningen. Den diskuteras för närvarande även inom europeiska och internationella standardiseringsorgan.)

Litteratur

Gavare, Rolf. 1988. Alphabetical Ordering in a Lexicological Perspective. I: *Studies in Computer-Aided Lexicology*: 63-102. Stockholm.