24. Språkstatistik – fortsättning

Som ni såg i de föregående avsnitten, underlättade det väldigt mycket att veta var klartextens ord börjar och slutar och därmed hur långa de är. Man måste förutsätta att kryptören också vet det och helt enkelt utesluter ordmellanrum och skriver meningarna i en enda lång rad. Möjligen finns det något som talar om när en mening börjar och slutar, men det ger inte forcören mycket hjälp. I fortsättningen antar vi därför att ordmellanrum inte används vid kryptering.

Men även med orden utpekade behövde ni mer än enkel pinnstatistik för språkets bokstäver för att knäcka kryptot. Ni behövde t ex veta några vanliga korta ord för att göra insteg. I det här avsnittet skall vi presentera mer språkstatistik och visa hur den kan användas för forcering.

I detta avsnitt finns data som anger hur ofta enskilda bokstäver samt kombinationer av två och tre bokstäver förekommer i svenska och engelska.

Det material som ligger till grund för de svenska statistikerna består av cirka 60 000 tecken modern tidningstext. De olika texterna har tagits från olika ämnesområden så att de resulterande statistikerna skall bli så neutrala som möjligt.

Motsvarande engelska material är ungefär lika stort som det svenska och har utvalts på samma sätt.

Monogramstatistik anger hur ofta enskilda bokstäver förekommer i en viss textmassa. Vi har helt uteslutit siffror, mellanslag och skiljetecken. Stora bokstäver har räknats som små.

Bigramstatistik anger hur ofta två bokstäver efter varandra förekommer i materialet. Textmassan har behandlats som en enda lång följd av bokstäver. Alla andra tecken har uteslutits.

Trigramstatistik anger hur ofta olika trebokstavskombinationer förekommer i textmassan. De har beräknats på samma sätt som bigramstatistik. De tal som anger hur ofta en bokstav eller bokstavskombination förekommer är procenttal.

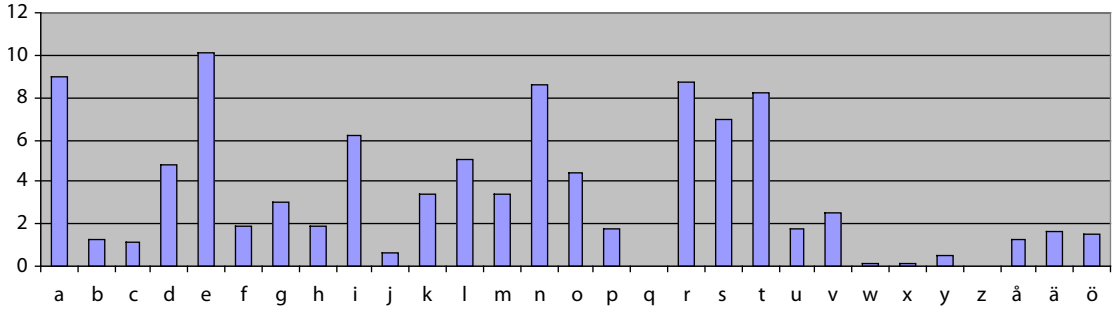
Monogramstatistik – svenska

I bokstavsordning

a	9,0
b	1,3
c	1,2
d	4,8
e	10,1
f	1,9
g	3,0
h	1,9
i	6,2
j	0,6
k	3,4
l	5,0
m	3,4
n	8,6
o	4,4
p	1,8
q	0,0
r	8,7
s	6,9
t	8,2
u	1,8
v	2,5
w	0,1
x	0,1
y	0,5
z	0,0
å	1,3
ä	1,7
ö	1,5

I ordning efter förekomst

e	10,1
a	9,0
r	8,7
n	8,6
t	8,2
s	6,9
i	6,2
l	5,0
d	4,8
o	4,4
m	3,4
k	3,4
g	3,0
v	2,5
h	1,9
f	1,9
p	1,8
u	1,8
ä	1,7
ö	1,5
å	1,3
b	1,3
c	1,2
j	0,6
y	0,5
x	0,1
w	0,1
z	0,0
q	0,0



Bigramstatistik – svenska

(Bigram som förekommer mer sällan än i 0,25 % av fallen har en tom ruta.)

Andra bokstav

Första bokstav	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z	å	ä	ö
a				0.6			0.4				0.3	0.6	0.5	1.7				1.5	0.6	1.1		0.6							
b					0.4																								
c								0.7			0.3																		
d	0.5				2.2													0.3	0.3										
e				0.6		0.3						0.6	0.3	2.5				2.2	0.7	1.2									
f																													0.7
g	0.5				0.8																								
h	0.5																												
i				0.4			0.6				0.4	0.6		1.3	0.3				0.9	0.4									
j																													
k	0.8				0.4										0.5					0.3									
l	0.7			0.3	0.5				0.7			0.9							0.3										
m	0.5				0.7				0.3				0.3																
n	1.0			1.2	0.5		0.8		0.7					0.4	0.4				0.9	0.6									
o			0.8									0.3	0.9	0.6				0.7											
p																0.3		0.3									0.3		
q																													
r	1.1			0.4	1.1				1.1		0.4		0.3	0.4	0.5				0.7	0.4									
s	0.5				0.4				0.5		0.8				0.6				0.4	1.5		0.3							
t	1.0				1.3				1.2						0.5			0.5	0.6	1.1									
u														0.4						0.3									
v	0.6				0.5				0.5																				
w																													
x																													
y																													
z																													
å														0.3															
ä														0.3				0.6											
ö																		0.8											

De 32 vanligaste bigrammen i svenska språket med sina procenttal:

en	2,5	et	1,2	ta	1,0	oc	0,8
er	2,2	nd	1,2	na	1,0	ng	0,8
de	2,2	ti	1,2	is	0,9	ör	0,8
an	1,7	ra	1,1	ll	0,9	ge	0,8
st	1,5	re	1,1	ns	0,9	rs	0,7
ar	1,5	at	1,1	om	0,9	li	0,7
in	1,3	tt	1,1	sk	0,8	me	0,7
te	1,3	ri	1,1	ka	0,8	es	0,7

Trigramstatistik – svenska

Det vanligaste trigrammet i svenska språket är 'för', tätt följt av 'och', 'nde' och 'and'. Vart och ett av dessa förekommer i en stor textmassa med frekvensen 0,6 %. De 60 vanligaste trigrammen i ordning efter hur vanliga de är:

för	des	men
och	var	ion
nde	med	han
and	ist	lan
ing	nin	und
ter	ers	sto
den	isk	ern
att	eri	ger
ade	ste	lle
gen	ten	ris
som	rna	örs
ens	are	ett
ill	lig	nga
det	ans	ent
ska	ena	upp
sta	ren	eno
til	ati	sam
era	nge	nte
rin	ver	man
der	rde	sen

Dessa trigram har en eller två vokaler. Det vanligaste trigrammet med bara konsonanter är 'str' som kommer på plats 61.

Här kommer en övning som ni kan göra med hjälp av trigramstatistiken och lite logiskt tänkande.

Övning 24A

De sex vanligaste trigrammen i svenska språket har krypterats med enkel substitution. Resultatet har blivit: TÄX, EUÅ, AZB, YIB, UÅZ, CUD i någon ordning. Vilka kryptogrupper betyder vad?

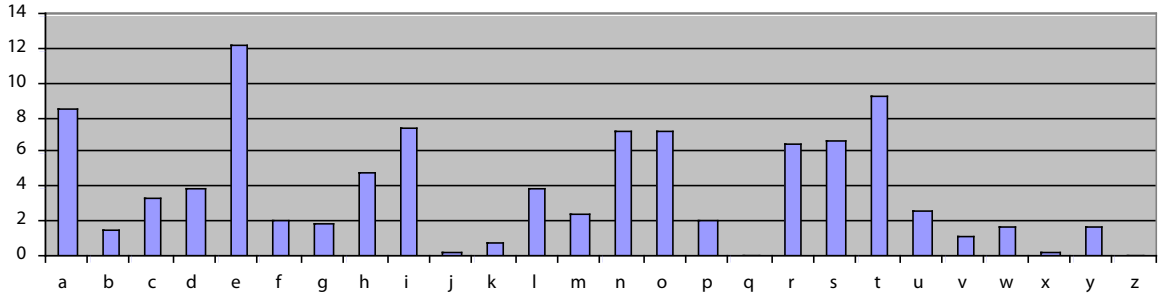
Monogramstatistik – engelska

I bokstavsordning

a	8,7
b	1,5
c	3,3
d	3,9
e	12,3
f	2,1
g	1,9
h	4,8
i	7,5
j	0,2
k	0,7
l	4,0
m	2,5
n	7,3
o	7,2
p	2,0
q	0,1
r	6,5
s	6,8
t	9,3
u	2,6
v	1,2
w	1,7
x	0,2
y	1,7
z	0,1

I ordning efter förekomst

e	12,3
t	9,3
a	8,7
i	7,5
n	7,3
o	7,2
s	6,8
r	6,5
h	4,8
l	4,0
d	3,9
c	3,3
u	2,6
m	2,5
f	2,1
p	2,0
g	1,9
w	1,7
y	1,7
b	1,5
v	1,2
k	0,7
x	0,2
j	0,2
q	0,1
z	0,1



Övning 24B

Jämför med svensk monogramstatistik. Vilka likheter och skillnader hittar ni?

Bigramstatistik – engelska

(Bigram som förekommer mer sällan än i 0,25 % av fallen har en tom ruta.)

första bokstav \ andra bokstav	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z
a			0.3	0.3					0.5			0.9	0.3	1.5				1.0	0.8	1.2					0.3	
b					0.4																					
c	0.5				0.6			0.3							0.6					0.3						
d	0.4				0.7				0.5						0.3					0.4						
e	1.1		0.7	1.1	0.4	0.3			0.4			0.4	0.5	1.3	0.3	0.3		1.9	1.3	0.7			0.3			
f															0.4					0.3						
g					0.3																					
h	1.0				2.2				0.5						0.4											
i			0.5	0.4	0.3							0.4		1.9	0.6			0.3	1.0	0.9						
j																										
k					0.3																					
l	0.6				0.6				0.6			0.4			0.3										0.3	
m	0.4				0.7				0.3						0.3											
n	0.7		0.5	0.9	0.6		0.8		0.4						0.4				0.5	1.3						
o						0.7						0.4	0.4	1.5				1.0	0.3	0.4	0.6		0.3			
p	0.3				0.4										0.3			0.3								
q																										
r	0.6				1.4				0.7						0.7				0.5	0.5						
s	0.7		0.3		0.8			0.3	0.7						0.5				0.5	1.3	0.3					
t	0.8				0.9			2.8	1.1						1.0			0.4	0.4	0.5						
u														0.3				0.4	0.3	0.3						
v					0.7																					
w	0.3				0.3			0.3	0.3																	
x																										
y																										
z																										

De 32 vanligaste bigrammen i engelska språket är med sina procenttal:

th	2,8	es	1,3	to	1,0	as	0,8
he	2,2	nt	1,3	is	1,0	se	0,8
in	1,9	en	1,3	ar	1,0	ng	0,8
er	1,9	at	1,2	ha	1,0	ta	0,8
on	1,5	ed	1,1	nd	0,9	de	0,7
an	1,5	ea	1,1	te	0,9	ec	0,7
re	1,4	ti	1,1	al	0,9	sa	0,7
st	1,3	or	1,0	it	0,9	et	0,7

Övning 24C

Jämför engelsk och svensk bigramstatistik. Vilka likheter och skillnader hittar ni?

Trigramstatistik – engelska

Det i särklass vanligaste trigrammet i engelska är 'the' som förekommer i 1,8 % av alla trigram. Det näst vanligaste är 'ing' som har procenttalet 0,6. Sedan följer i tur och ordning 'ent', 'and', 'ion', 'tha' och 'tio', som har sannolikheter från 0,6 till 0,4.

De 36 vanligaste trigrammen i engelska är, ordnande efter sina procenttal:

the	int	ear	nce
ing	ter	ers	men
ent	ere	eth	res
and	her	ons	ont
ion	est	sin	rea
tha	for	dth	are
tio	ati	sta	eco
nth	ver	con	sth
hat	tth	ist	ith

Övning 24D

Några av de vanligaste trigrammen är egna ord. Vilka är det? De andra trigrammen ingår som delar i ord eller bildar övergång mellan två ord som kommer efter varandra. För vart och ett av dessa trigram, försök att finna ord som innehåller just det trigram som ni gått ut ifrån. Exempel: 'ter' ingår i 'winter'. Vem av er hittar riktiga engelska ord till flest trigram på fem minuter?

