

STATISTISKA METODER I SPRÅKTEKNOLOGI

VIGGO KANN

FÖRELÄSNING 1: SPRÅKSTATISTIK

- VAD KAN MAN GÖRA STATISTIK PÅ?
- VAD HAR MAN STATISTIKEN TILL?
- VILKA KÄLLOR TILL STATISTIK FINNS?
- HUR TAR MAN FRAM STATISTIK?

ASCII	Chr	Frekvens	Procent	90	Z	22190	0.01	194	Å	1	0.00
10		4437349	1.42	91	[	16813	0.01	195	Ä	19	0.00
32		45789122	14.69	92	\	9	0.00	196	Å	45947	0.01
33	!	29431	0.01	93	]	16814	0.01	197	Ä	44341	0.01
34	"	373077	0.12	94	^	2567691	0.82	198	Æ	72	0.00
35	#	96	0.00	95	-	855	0.00	199	Ç	345	0.00
36	\$	474	0.00	96	~	109	0.00	200	È	197	0.00
37	%	1556	0.00	97	a	22474000	7.21	201	É	1057	0.00
38	&	14010	0.00	98	b	3079887	0.99	202	Ê	29	0.00
39	'	17266	0.01	99	c	2998176	0.96	203	Ë	6	0.00
40	(	201461	0.06	100	d	9809271	3.15	204	Ì	1	0.00
41	)	228414	0.07	101	e	23849906	7.65	205	Í	38	0.00
42	*	5845	0.00	102	f	4581350	1.47	206	Î	7	0.00
43	+	6414	0.00	103	g	7441595	2.39	207	Ï	1	0.00
44	,	1891414	0.61	104	h	4438644	1.42	209	Ñ	2	0.00
45	-	929090	0.30	105	i	13703278	4.40	210	Ò	4	0.00
46	.	3107876	1.00	106	j	1514856	0.49	211	Ó	52	0.00
47	/	54673	0.02	107	k	7650120	2.45	212	Ô	4	0.00
48	0	812583	0.26	108	l	12466840	4.00	213	Õ	15	0.00
49	1	843180	0.27	109	m	7766643	2.49	214	Ö	65934	0.02
50	2	495124	0.16	110	n	20778613	6.67	216	Ø	152	0.00
51	3	385983	0.12	111	o	10282367	3.30	217	Ù	25	0.00
52	4	339270	0.11	112	p	4253071	1.36	218	Ú	16	0.00
53	5	399431	0.13	113	q	25890	0.01	219	Û	18	0.00
54	6	285869	0.09	114	r	21058002	6.76	220	Ü	467	0.00
55	7	259113	0.08	115	s	15231656	4.89	223	Ž	28	0.00
56	8	265250	0.09	116	t	20662695	6.63	224	à	1259	0.00
57	9	452335	0.15	117	u	4115639	1.32	225	á	2392	0.00
58	:	298514	0.10	118	v	5227239	1.68	226	â	246	0.00
59	;	24349	0.01	119	w	121644	0.04	227	ã	61	0.00
60	<	33649	0.01	120	x	278843	0.09	228	ä	4665240	1.50
61	=	1117	0.00	121	y	1499017	0.48	229	å	3794299	1.22
62	>	33542	0.01	122	z	90005	0.03	230	æ	1054	0.00
63	?	80684	0.03	123	{	7	0.00	231	ç	822	0.00
64	@	433	0.00	124		40	0.00	232	è	3373	0.00
65	A	542114	0.17	125	}	5	0.00	233	é	72812	0.02
66	B	444854	0.14	126	~	169	0.00	234	ê	302	0.00
67	C	223041	0.07	161	i	8	0.00	235	ë	659	0.00
68	D	785721	0.25	163	ë	34	0.00	236	ì	23	0.00
69	E	534662	0.17	164	ñ	11233	0.00	237	í	625	0.00
70	F	400951	0.13	165	¥	1	0.00	238	î	151	0.00
71	G	270642	0.09	167	§	184	0.00	239	ï	114	0.00
72	H	412108	0.13	168	"	187	0.00	241	ñ	187	0.00
73	I	385539	0.12	169	°	4686	0.00	242	ò	63	0.00
74	J	263838	0.08	170	°	150	0.00	243	ó	1285	0.00
75	K	341125	0.11	171	«	5	0.00	244	ô	450	0.00
76	L	373768	0.12	172	¬	57	0.00	245	ö	27	0.00
77	M	587433	0.19	174	ø	4419	0.00	246	ö	3538513	1.14
78	N	431895	0.14	175	-	2	0.00	247	÷	41	0.00
79	O	286775	0.09	176	°	68	0.00	248	°	1840	0.00
80	P	345619	0.11	177	±	194	0.00	249	ù	10	0.00
81	Q	8095	0.00	180	'	8997	0.00	250	ú	199	0.00
82	R	342978	0.11	181	µ	2	0.00	251	û	32	0.00
83	S	1019792	0.33	182	¶	18	0.00	252	ü	8277	0.00
84	T	485512	0.16	183	·	12972	0.00	253	ý	1	0.00
85	U	231490	0.07	184	.	64	0.00	255	y	26	0.00
86	V	334814	0.11	187	»	41	0.00				
87	W	102840	0.03	191	¿	2	0.00				
88	X	7662	0.00	192	À	46	0.00				
89	Y	37004	0.01	193	Á	57	0.00				

tot = 311,618,803

Tabell 3.2 Förekomst och frekvens av enskilda tecken i textfilen svd9596.txt.

DE FEM VANLIGASTE orden i de fem spexen är *jag* (1119 förekomster), *det* (985), *är* (942), *och* (720) och *att* (706). De hundra närmast vanligaste orden är i tur och ordning *en*, *inte*, *du*, *i*, *som*, *på*, *har*, *för*, *vi*, *så*, *men*, *den*, *här*, *med*, *ni*, *till*, *nu*, *ska*, *ju*, *om*, *kan*, *han*, *var*, *vad*, *ett*, *min*, *mej*, *av*, *där*, *får*, *ja*, *man*, *nej*, *då*, *måste*, *bara*, *vara*, *er*, *dom*, *när*, *få*, *vill*, *väl*, *hur*, *din*, *kommer*, *dej*, *ha*, *skulle*, *mitt*, *blir*, *Ivan*, *bli*, *se*, *ut*, *in*, *publiken*, *gör*, *oss*, *från*, *ta*, *vet*, *upp*, *gå*, *än*, *sej*, *alla*, *bra*, *går*, *va*, *göra*, *kanske*, *ingen*, *tror*, *sen*, *nog*, *lite*, *kom*, *allt*, *tar*, *mot*, *just*, *vår*, *någon*, *nåt*, *hon*, *e*, *åh*, *ditt*, *vilken*, *ser*, *rätt*, *något*, *lilla*, *eller*, *honom*, *Wagner*, *sin*, *låt*, *efter* (38 förekomster).

De 12 vanligaste orden i svenskt skriftspråk enligt NFO 1	
ORD	ANTAL (på 1 million löpord)
1. <i>och</i> koordinator	30856
2. <i>i</i> prep	29461
3. <i>en</i> artikel	18184
4. <i>är</i> kopula	15506
5. <i>av</i> prep	15262
6. <i>det</i> pron	13991
7. <i>på</i> prep	13932
8. <i>som</i> pron	13522
9. <i>att</i> infinitivmärke	12534
10. <i>för</i> prep	11807
11. <i>med</i> prep	11700
12. <i>att</i> konjunktion	11232

Längd	Frekv	Proc	Kumul
1	2330973	4.9%	4.9
2	5682243	12.0	16.9
3	11091917	23.4	40.3
4	5194141	11.0	51.2
5	5160818	10.9	62.1
6	4413182	9.3	71.4
7	3200004	6.7	78.2
8	2642731	5.6	83.7
9	2156242	4.5	88.3
10	1631150	3.4	91.7
11	1115255	2.4	94.1
12	770439	1.6	95.7
13	530572	1.1	96.8
14	406597	0.9	97.7
15	322407	0.7	98.3
16	224162	0.5	98.8
17	164475	0.3	99.2
18	130690	0.3	99.4
19	84100	0.2	99.6
20	57714	0.1	99.7
21	44227	0.1	99.8
22	24383	0.1	99.9
23	15947	0.0	99.9
24	22622	0.0	100.0
25	5573	0.0	100.0

Utdrag ur SUC:

```

("<s>>"          <81>)
("<Redan>"       <82> (AB "redan"))
("<en>"          <83> (DT UTR SIN IND "en"))
("<f|rv{ntan}>"  <84> (NN UTR SIN IND NOM "f|rv{ntan}"))
("<att>"         <85> (SN "att"))
("<PTK>"         <86> (PM NOM "PTK"))
("<s{ger}>"      <87> (VB PRS AKT "s{ga}"))
("<nej>"         <88> (IN "nej"))
("<kan>"         <89> (VB PRS AKT "kunna"))
("<leda>"        <90> (VB INF AKT "leda"))
("<till>"        <91> (PP "till"))
("<att>"         <92> (SN "att"))
("<andra>"       <93>
(JJ POS UTR/NEU PLU IND/DEF NOM "annan"))
("<ocks}>"      <94> (AB "ocks}"))
("<s{ger}>"      <95> (VB PRS AKT "s{ga}"))
("<nej>"         <96> (IN "nej"))
("<.>"          <97> (DL MAD ".")
("<</s>>"       <98>)

```

Utdrag ur BNC:

```

<head type=MAIN>
<s n=001>
  <w NN1>FACTSHEET <w DTQ>WHAT <w VBZ>IS <w NN1>AIDS<c PUN>?
</head>
<p>
<s n=002>
<hi r=bo> <w NN1>AIDS <c PUL>(<w NP0>Acquired <w AJ0>Immune
  <w NN1-NP0>Deficiency <w NP0>Syndrome<c PUR>) </hi>
  <w VBZ>is <w AT0>a <w NN1>condition <w VVN>caused <w PRP>by
  <w AT0>a <w NN1>virus <w VVD-VVN>called <w NP0>HIV
  <c PUL>(<w AJ0-NN1>Human <w NN1-NP0>Immuno <w NP0>Deficiency
  <w NP0>Virus<c PUR>)<c PUN>.
<s n=003>
  <w DT0>This <w NN1>virus <w VVZ>affects <w AT0>the
  <w NN1>body<w POS>'s <w NN1>defence <w NN1>system
  <w CJS>so that <w PNP>it <w VM0>can<w XX0>not
  <w VVI>fight <w NN1>infection<c PUN>.
</p>

```

KORPUS

SUC

STOCKHOLM-UMEA-KORPUS (1997)
1 MILJON ORD, BALANSERAD KORPUS
TILLGÄNGLIG FÖR FORSKNING, CD

NFO

NUSVENSK FREKVENSORDBOK (1970)
1 MILJON ORD, TIDNINGSTEXT
4 BÖCKER, FINNS PÅ STADSBIBLIOTEKET

SPRÅKBANKEN

PÅ SPRÅKDATA VID GÖTEBORGS UNIV.
DIVERSE TEXTER, SAOL, SAOB
KONKORDANSER

PROJEKT RYNEBERG

LYSATOR VID LINKÖPINGS UNIV.
LITTERÄRA TEXTER, FRITT TILLGÄNGLIGT

BNC

BRITISH NATIONAL CORPUS (1995)
100 MILJONER ORD, BALANSERAD KORPUS
KAN KÖPAS PÅ CD/pkg/corpus/1.0

PAROLE

spraakdata.gu.se 19 MILJONER ORD

TOKENISERING

UPPDELNING AV TEXT I ENSKILDA ORD.

EXEMPEL:

IDAG 28/11 HAND-
LADE FÖRELÄSNINGEN
BL. A. OM SUC-
KORPUSEN.

IDAG
28/11
HANDLADE
FÖRELÄSNINGEN
BL. A.
OM
SUC-KORPUSEN

PROBLEM:

- SKILJETECKEN BÖR VARA EGNA ORD, MEN PUNKTER HÖR TILL FÖRKORTNINGARNA.
- BINDESTRECK SIKT PÅ RADER.
- TÄNKSTRECK UTAN MELLANRUM RUNT.

ENKEL TOKENISERING I UNIXMILJÖ:

tr ' ' '\012' < infil > fil1

ERSÄTTER ALLA MELLANSLAG MED RADBYTEN

tr -s '\012' < fil1 > fil2

ERSÄTTER FLERA RADBYTEN I RAD MED ETT ENDA

EXTRAHERA ORDLISTA UR TOKENISERAD TEXT:

tr '[A-Z][a-z]' '[a-z]' < fil2 > fil3

OMVANDLAR TILL SMÅ BOKSTÄVER

sed 's/[A-Z][a-z-g0-9]\$//>' < fil3 > fil4

TAR BORT AVSLUTANDE SKILJETECKEN

sort -d fil4 | uniq -c > frekvensordlista

SORTERAR OCH BERÄKNAR ORDFREKVENSER

UNIXPROGRAM FÖR KORPUSARBETE

grep, egrep

RADVIS SÖKNING

sort

SORTERING

uniq

SAMMANSLAGNING AV LIKA RADER

tr

BOKSTAVSVIS UTBYTE

wc

RÄKNING AV RADER, ORD, TECKEN

cut

PLOCKA UT VISST FÄLT FRÅN VARJE RAD

paste

SLÅ IHOP FILER RADVIS

join

RELATIONSDATABASESOPERATIONEN JOIN

comm

KÖLL AV LIKADANA RADER I SORTERADE FILER

cat

SAMMANSLAGNING AV FILER

tail

UTPLOCKNING AV SLUTET AV FIL

sed

STRÖMEDITOR