Bengt Beckman
Vitklövervägen 38
16360 SPÅNGA

UTNYTTJANDE AV ORDKLASSER FÖR FÖRFATTARBESTÄMNING

Hösten 1974 dök det i Paris upp en sensationell skrift i vilken 1965 års nobelpristagare Michail Sjolochov anklagades för plagiat. Studien hade skrivits av en senare avliden sovjetisk kritiker D, och förordet hade skrivits av Alexander Solzjenitsyn, som helt stödde slutsatsen: det mesta av Stilla flyter Don har inte skrivits av Michail Sjolochov utan av en annan kosackförfattare, Fedor Krjukov.

Detta är bakgrunden till att det hösten 1975 bildades ett svenskt-norskt team som tog sig an frågan: "Vem har skrivit Stilla flyter Don?" Deltagare var Sven Gustavsson och Bengt Beckman från Sverige och Geir Kjetsaa och Steinar Gil från Norge. Avsikten med projektet var dels att lösa författarproblemet, dels att testa kvantitativa metoder för stilanalys och författarskapsundersökningar.

Undersökningen är nu - efter några avbrott - slutförd och resultaten redovisas i bokform.[1] Föredraget behandlar en av de studier som gjorts.

Studien utgör en utvidgning av en metod som beskrivits vid två tidigare tillfällen, dels vid ett låtsat författarbestämningsförfarande på material av de sovjetiska författarna K Simonov och K Paustovskij, dels vid de preliminära undersökningarna i det mera spektakulära fallet med Krjukov och Sjolochov som möjliga författare till Stilla flyter Don.

Det kriterium som användes i det första fallet var ordklassfördelning i meningsbörjan och meningsslut. I Geir Kjetsaas pilotstudie "Storms on the Quiet Don" undersöktes Sjolochov, Krjukov och "Stilla flyter Don" ur bl a denna aspekt. Resultaten, som klart talar för Sjolochov, har presenterats i andra sammanhang och skall ej beröras här. Emellertid bekräftade undersökningen ordklasskriteriets användbarhet. Jag har därför i den större, datorstödda undersökningen i samma ämne, vars resultat nu föreligger, använt ordklasskombinationer totalt och inom hela meningar som kriterium.

---

[1] Geir Kjetsaa, Sven Gustavsson, Bengt Beckman, Steinar Gil.
The Authorship of The Quiet Don. Slavica Norvegica II. Oslo 1983.

Utnyttjande av ordklasser for författarbestämning
Bengt Beckman
Proceedings of NODALIDA 1983, pages 21-28

Ordklassindelningen i nämnda mindre undersökningar gjordes manuellt. Detta är - frånsett det besvärliga i arbetet - otillfredsställande ur den synpunkten att objektivitet och konsekvens ej kan garanteras. När Sven Gustavsson tog initiativet till projektet "Vem har skrivit Stilla flyter Don?" fanns ett automatiskt analyssystem uppbyggt vid Uppsala Datacentral under ledning av Anna Sågvall Hein. Systemet delar in orden i morfologiska klasser efter böjningsändelser och andra formella kännetecken. Detta resulterar i en modifierad ordklassindelning, som får fördelen att vara konsekvent, vilket ju i detta sammanhang är betydligt viktigare än att den är mänskligt rätt. Systemet har modifierats och utökats för att passa den aktuella undersökningen. Sålunda har klassindelningen genom att tillvarata underklasser ökats till att omfatta 25 klasser. Systemet utför en klassbestämning av varje ord och översätter varje mening till en svit klassbetecknande bokstäver. (Fig 1, Fig 2)

Sammanlagt har 176 500 ord fördelade på 12 400 meningar skrivits in till analyssystemet via terminaler. Till systemet hör ett stamlexikon. Varje nytt ord som tillförs lexikonet måste förses med en serie kriterier för att ordet när det nästa gång dyker upp i någon annan böjningsform skall kunna identifieras automatiskt. Denna uppdatering av lexikonet har blivit en omfattande uppgift. Som tidigare nämnts modifierades systemet och många tilläggsprogram skrevs. Detta arbete skedde vid Uppsala Datacentral och - på senare tid - vid Uppsala universitet, Centrum för datorlingvistik.

På materialet i klassbetecknad form har gjorts ett flertal statistiska undersökningar. Alla resultat tyder på samma sak. Ordklassfördelningen skiljer sig från Krjukovs men är lik Sjolochovs. Som exempel har tagits 4-klasskombinationer. I Fig 3 anges under H_0 de förväntade frekvenserna framräknade ur äkta Krjukov-material och under H_1 motsvarande ur Sjolochov-material. Under X_1 återfinns de i det omstridda verket observerade frekvenserna.

Den svaghet som vidlåder de flesta metoder är, att också det omstridda materialet måste vara ganska stort för att duga till statistiska jämförelser. Vad man eftersträvar är ett test, genom vilket man kan utläsa trolig författartillhörighet på en mindre mängd omstritt material, t ex 20-30 meningar. Detta test bör vara så beskaffat, att man - mening för mening - tillvaratar alla i meningen ingående syntaktiska två-kombinationer och jämför med sannolikheter framräknade ur resp författares oomstridda verk. Genom att multiplicera sannolikheterna i varje mening får man ett mått på hur bra den är som - i detta fallet - Sjolochovmening resp Krjukovmening. Två test enligt denna modell har prövats på det aktuella materialet.

Fig 1. Klassindelning och klassbeteckningar.

A	Nominatives of animates	Basic class 1
B	Other cases of animates	
C	Nominatives of inanimates	
D	Other cases of inanimates	
E	Nominatives	Basic class 2
F	Other forms	
G	Long forms, nominative	Basic class 3
H	Long forms, other cases	
I	Short forms, comparative forms like jasnee	
J	Nominatives of kto, čto, my, vy, ja, ty	Basic class 4
K	Other forms of pronouns mentioned in J	
L	Other members of basic class 4	
M	All forms	Basic class 5
N	Pure prepositions like bez, bezo, v, vo	Basic class 6
O	Prepositions which can also be adverbs, for example vblizi, vdol', vnutri	
P	Pure conjunctions like a, i, ili	
Q	Conjunctions which can also be adverbs, for example da, odnako, poka	
R	Other members of basic class 6	
S	Infinitives	Basic class 7
T	Gerunds	
U	Participles, long forms, nominative	
V	Participles, long forms, other cases	
W	Participles, short forms	
X	Other forms belonging to basic class 7	
Y	All punctuation marks other than sentence-dividing marks	

Fig 2. Meningar översatta till klassbetecknande bokstäver samt de ursprungliga meningarna.

```
 6  G C Y N H D C
15  C N F C X N C N D
45  G G C N H N C H D Y P R C Y G C D Y G
    G C V D C P I Y U N D H D C D
70  N C Y N D H D Y G C Y G C Y U H D G Y
    G C Y C N D Y N F Y U H D C
75  N D Y G C C
```

1. Melechovskij dvor - na samom kraju chutora.
2. Vorotca so skotin'ego baza vedut na sever k Donu.
3. Krutoj vos'misažennyj spusk meždu zamšelych v prozeleni melovych glyb, i vot bèreg: perlamutrovaja rossyp' rakušek, seraja izlomistaja kajma nacelovannoj volnami gal'ki i dal'še - perekipajuščee pod vetrom voronenoj rjab'ju strem'ja Dona.
4. Na vostok, za krasnotalom gumennych pletnej, - getmanskij šljach, polynnaja prosed', istoptannyj konskimi kopytami buryj, živuščoj pridorožnik, časovenka na razvilke; za nej - zadernutaja tekučim marevom step'.
5. S juga - melovaja chrebtina gory.

Fig 3. Tetragramfrekvenser.
H_0=förväntade frekvenser framräknade ur Krjukov-material.
H_1=förväntade frekvenser framräknade ur Sjolochov-material.
X_1=obseverade frekvenser i Stilla flyter Don (Tichij Don = TD)

TD total (n = 44,408)

TETRA	H_0 expected	H_1 expected	X_1 observed
CYXN	38.63	99.03	104
DYNH	83.04	45.74	50
DYXN	50.18	127.45	195
GYGC	81.27	19.54	23
GYGY	91.92	19.54	24
HDYN	71.50	48.40	50
HYHD	95.92	23.09	25
NCYX	45.30	92.37	76
NDYX	83.04	124.34	149
NFDY	63.94	35.52	38
NHDX	51.07	72.83	67
RXND	39.52	58.62	29
XNCY	59.95	133.22	111
XNDY	115.90	166.53	166
YGYG	82.15	18.65	26
YNHD	96.81	64.39	96
YXNC	41.74	129.23	118
YXND	64.84	130.12	196
YXYX	69.72	35.97	42

Test I. Vi antar att en text ordklassmässigt är en Markovkedja, d v s att ett ords ordklasstillhörighet har en sannolikhetsfördelning som enbart beror av föregående ords ordklass. Vi talar om övergångssannolikheter p(i, j), d v s sannolikheten att ett ord i ordklass i följs av ett ord i ordklass j. Dessa övergångssannolikheter är positionsberoende. Speciellt stort är positionsberoendet i början och slutet av meningen. Dessutom inför vi sannolikheten P_i att en mening börjar med ordklass i. Alla dessa sannolikheter bestäms genom statistik som upprättats på okänt material. Vi får då olika uppsättningar sannolikheter för olika författare. Våra testvariabler blir således (meningen antas vara $x_1\ x_2\ x_3 \ldots x_n$)

$$P_S = P^S_{x_1} \prod_{i=1}^{N-1} P^S_i\,(x_i, x_{i+1})$$

$$P_k = P^K_{x_1} \prod_{i=1}^{N-1} P^k_i\,(x_i, x_{i+1}) \qquad \text{TEST I}$$

Test II. Ett problem vid testvalet är: hur bra är approximationen att vi har Markovkedjor? I varje fall inte perfekt. Som alternativt test har därför använts följande. Låt $q^s_k(i, j)$ vara sannolikheten att bigrammet på plats k och k+1 är ett ord i ordklassen i och ett i ordklassen j (hos författaren S).

$$P_S = \prod_{i=1}^{N-1} q^S_i\,(x_i, x_{i+1})$$

$$P_K = \prod_{i=1}^{N-1} q^K_i\,(x_i, x_{i+1}) \qquad \text{TEST II}$$

Detta test (Test II) ger högre värden där författaren använt för sig vanliga ordklasser. Något överraskande ger test II bättre resultat än test I när det prövats på samma S- och K-texter.

Testen applicerades på 10-meningsavsnitt ur Stilla flyter Don (Tichij Don= TD). Resultaten exemplifieras av fig 4.

Fig 4.

TD,1

Number of sentences		Word number	Number of words	Test I Sentences of type S	Test I Sentences of type K	Test II Sentences of type S	Test II Sentences of type K	I Predominantly	II Predominantly
1	10	1-123	123	6	4	8	2	S	S
2	10	801-914	114	7	3	10	0	S	S
3	10	1384-1510	127	2	8	7	3	K	S
4	10	2001-2079	79	9	1	9	1	S	S
5	10	2573-2661	89	6	4	8	2	S	S
6	10	3148-3212	65	5	5	7	3	X	S
7	10	3671-3855	185	9	1	9	1	S	S
8	10	4398-4498	101	5	5	7	3	X	S
9	10	5012-5092	81	5	5	8	2	X	S
10	10	5630-5744	145	7	3	8	2	S	S
11	10	6415-6565	151	6	4	5	5	S	X
12	10	6916-6987	72	7	3	9	1	S	S
13	10	7363-7479	117	7	3	8	2	S	S
14	10	7991-8075	85	4	6	7	3	K	S
15	10	8654-8777	124	8	2	9	1	S	S
16	10	9231-9305	74	4	6	8	2	K	S
17	10	9911-10115	204	6	4	6	4	S	S
18	10	10915-11063	149	7	3	7	3	S	S
19	10	11573-11695	123	6	4	10	0	S	S
20	10	12320-12384	65	5	5	8	2	X	S

Sammanfattning av resultat

10-meningsavsnitt.

Stickprov ur äkta Sjolochov- och Krjukov-texter:

	Test I			Test II		
	S	K	X	S	K	X
Äkta S	12	5	6	22	1	0
Äkta K	0	22	1	3	17	3

Stickprov ur Stilla flyter Don (Tichij Don):

	Test I			Test II		
	S	K	X	S	K	X
TD 1	13	3	4	19	0	1
TD 2	13	4	3	19	0	1
TD 3	15	6	3	22	0	2
TDtot	41	13	10	60	0	4

Under S noteras antal 10-meningsavsnitt med dominans av S-meningar.
Under K noteras antal 10-meningsavsnitt med dominans av K-meningar.
Under X noteras antal 10-meningsavsnitt där 5 meningar är av S-typ och 5 är av K-typ.

Anm. Flera skäl kan tänkas till testets goda diskriminerande förmåga, t ex:
- de 25 klasserna är en lämplig abstraktionsnivå för ord
- kombinationer av två klasser är en nivå på vilken individuella vanor klart framträder
- meningskonstruktion, i det avseende som testet avspeglar, är en djupt rotad vana, som inte förändras vare sig av ämnesområde eller författarens utveckling.