Lennart Lönngren

DEN RYSKA TEXTKORPUSEN I UPPSALA. NÅGRA PRELIMINÄRA RESULTAT.

Vid Uppsala universitet pågår ett projekt som syftar till att framställa en modern rysk textkorpus samt utföra vissa bearbetningar av denna. I projektet medverkar, förutom jag själv, fil. lic. Ludmila Ferm. Textmassan omfattar en miljon löpord och är jämnt fördelad mellan fack- och skönprosa. En första presentation av korpusen, framför allt avseende principer för val av texter, publicerades i *Slovo* 36 (1988), s. 77-83.

Till dags dato (april 1990) är samtliga texter inmatade i datorn, och arbetet med automatisk textbearbetning av hela textmassan har påbörjats. Emellertid genomfördes redan efter etappmålet en halv miljon ord, vilket uppnåddes våren 1989, vissa bearbetningar och undersökningar; dessa har förhoppningsvis en viss relevans, tack vare vår medvetna strävan att korpusen redan på denna nivå skulle vara mångsidigt sammansatt, med samma struktur och generella textfördelning som på enmiljonnivån. I denna artikel skall jag redogöra för några preliminära resultat av dessa undersökningar, speciellt några iakttagelser som rör lexikaliska skillnader mellan fack- och skönprosa.

Trots att det existerar avancerade metoder för statistiska korpusbeskrivningar, se t.ex. Arapov 1988, kommer jag här att använda mig av mycket elementära begrepp och beräkningar. Det kommer till stor del att röra sig om punktvisa, konkreta observationer av kvantitativa förhållanden med mycket tydliga tendenser. Om alltså min metodologi i detta sammanhang är ganska primitiv vill jag å andra sidan hävda att de faktiska lexikaliska värden som här presenteras på grundval av den

nyproducerade korpusen – även på halvmiljonnivån – sannolikt har en större aktualitet och tillförlitlighet än tidigare kvantitativa uppgifter om ryskan. Dessa har huvudsakligen – så t.ex. hos Arapov – haft som källa den korpus som ligger till grund för Zasorinas frekvensordbok från 1975, *Častotnyj slovar' russkogo jazyka* (nedan *ČS*). Beträffande språkets aktualitet kan nämnas att medan Zasorinas texter sträcker sig ända från seklets början – bl.a. finns Lenin och Gorkij representerade – så tillämpas i Uppsalakorpusen fasta bakre tidsgränser för textvalet, nämligen 1960 för skönlitterära texter och 1985 för facktexter; det stora gapet mellan dessa båda tidsgränser är betingat av att fackprosa inte minst lexikaliskt åldras mycket snabbare än skönprosa. Påståendet att Uppsalakorpusen är mer tillförlitlig baserar jag på att den är mera strikt uppbyggd än Zasorinas. Stora ansträngningar har gjorts för att textmassan skall bli så representativ som möjligt – naturligtvis inom den givna ramen, dvs det kodifierade ryska skriftspråket. För fackprosans del innebär detta att vi har försökt täcka in alla kunskapsområden (teman) och ge varje tema en rätt viktning. Beträffande de skönlitterära texterna fördelar sig dessa på 40 författare, varvid mer betydande författare är representerade med större textmassa än mindre betydande. En sammanställning av teman/författare samt fördelning på sampler ges i appendix 1.

En uppfattning om hur väl vi har lyckats att göra korpusen väl sammansatt och representativ kan vi få genom att jämföra den första halvmiljonen av korpusen med den andra. Ordformer i de övre frekvensskikten bör då hamna på ungefär samma inbördes plats (och ha ungefär samma frekvens). Det visar sig att så också är fallet. För de 60 mest frekventa lexemen är kvoten mellan det lägre och det högre frekvenstalet aldrig lägre än 0,8 (detta "bottenvärde" representeras av *možet*, med frekvensen 659 i undre och 822 i övre halvan; ordformens rang är 50 resp. 43).

Av intresse i detta sammanhang kan det vara att fixera vilket som är ryskans mest frekventa lexem. Enligt *ČS* är det prepositionen *v* (inkl. formen *vo*), vilken i denna frekvensordbok segrar över konjunktionen *i* med c:a 6.500 förekomster, men den bearbetning vi nu tagit fram pekar entydigt på att *i* är det mest frekventa lexemet; i hela korpusen förekommer konjunktionen *i* 38.096 gånger medan prepositionen *v* förekommer 33.063 gånger (varav formen *vo* 1.196 gånger). Om vi ser uteslutande till den halva miljonen facktext, däremot, har *v/vo* en knapp ledning (18.796 mot 17.301).

Korpusen är uppbyggd av sampler om 5.000 löpord; varje sample kan i sin tur bestå av en eller flera texter. Genomsnittligt består ett sample av knappt tre texter (fler i fackprosa, färre i skönprosa). I sin helhet består korpusen således av 200 sampler; antalet texter är nära 600. Uppdelningen i sampler används bl a för att räkna fram det mycket viktiga spridningsvärdet hos lexikaliska enheter, varom mera nedan.

De textbearbetningar vi hittills har utfört innebär bl.a. att vi har tillgång till en lista över alla ordformer, dels över den totala textmängden på en miljon löpord, dels separat över 500.000 ord fackprosa och lika mycket skönprosa. Ordformslistorna kan vara försedda med olika slags information och vara sorterade efter olika kolumner. En vanlig utskrift utgörs t.ex. av en alfabetiskt sorterad lista med angivande av total frekvens och beläggställen i korpusen; den senare informationen ges genom hänvisning till ordformernas löpnummer vilken i sin tur återfinns i en speciell mening-för-mening-utskrift av textfilen. Man kan också erhålla ordformslistan sorterad primärt efter frekvens, sekundärt alfabetiskt med angivande av dels (total) frekvens, dels "rang", varvid rangvärdet sammanfaller med ordningsnumret för den sista ordformen i varje frekvensområde (dvs alla ordformer med samma frekvens får samma rang). Sådana frekvenslistor föreligger för a) hela korpusen; b) separat för fack- och skönprosa; c) separat för undre och övre hälften av

korpusen (för ovan nämnda kontrolländamål).

Vid såväl inmatning av data som utskrift av resultatfiler används ett speciellt translittereringssystem, vilket så nära som möjligt ansluter till den internationella translittereringen; de avvikelser som finns beror på att vi dels inte kan använda några diakritiska tecken, dels har bestämt oss för att tillämpa ett-till-ett-förhållande mellan kyrilliska och latinska bokstäver.

Det finns möjlighet att med samma programpaket ("Textpack", framställt på Centrum för datorlingvistik i Uppsala av Valentina Rosén och Margareta Sjöberg) få fram en lista över grundformer (lemman) och deras frekvenser, men detta kan inte göras automatiskt. Det krävs ett interaktivt steg, eftersom grundformen måste anges manuellt för varje ordform. Detta innebär också att vissa icke-triviala lingvistiska beslut måste fattas, vilket naturligtvis i sin tur medför att resultaten blir något osäkra och icke helt jämförbara med andra undersökningar. Vi har emellertid i vårt projekt ännu inte kommit därhän; vi rör oss än så länge helt och hållet på ordformsnivån.

Inte ens i en stor korpus sammansatt av många texter är den framräknade totalfrekvensen tillräcklig som indikator på hur "viktigt" eller "centralt" ett ord är i språket i fråga. Ett ord kan nämligen erhålla en förvånansvärt hög frekvens inom ramen för en enda text. Så är t.ex. i vår korpus fallet med det ovanliga namnet *Bim*, som fått frekvensen 91 och därmed kommit upp i ett mycket centralt frekvensskikt. Det är allmänt erkänt att man erhåller ett bättre värde om man också tar hänsyn till spridningen. Om man måste gå efter enbart antingen frekvens eller spridning är sannolikt den senare faktorn mer rättvisande (såsom man har gjort i en rysk frekvenslista från 1968), förutsatt att man har ett tillräckligt antal olika deltexter. Bäst är emellertid att göra en sammanvägning mellan totalfrekvens och spridning enligt någon lämplig formel. Jag har valt att använda mig av samma formel som i

Sture Allén *et al.*, *Nusvensk frekvensordbok* (1970, band 1, sid XXVII):

$$D = 1 - \frac{s}{m\sqrt{n-1}}$$

där m är medelvärdet, n antalet deltexter och s standardavvikelsen. D står för "dispersion", dvs spridning. Formelns effekt är att modifiera utgångsfrekvensen med en faktor som varierar mellan ett och noll, varvid för värdet ett krävs att ordet har samma frekvens i alla deltexter (vi förutsätter att dessa alla är lika stora). Den andra ytterligheten är att ordet bara förekommer i en deltext, varvid värdet blir noll. Det framräknade värdet kan lämpligen kallas modifierad frekvens.

Vid Centrum för datorlingvistik i Uppsala har nyligen (av Bengt Dahlqvist) framställts ett program varmed vi har beräknat den modifierade frekvensen hos alla ord med minst frekvensen 2 i den första halvmiljondelen av korpusen. Detta innebär att dessa ordformer har utvärderats med avseende på spridningen. En stor del av dem har härvid visat sig inte "förtjäna" att stå kvar så högt som totalfrekvensen utvisar utan fått ett reducerat frekvenstal och "kanat ned" i listan. Det ovannämnda exemplet *Bim* har sålunda erhållit den modifierade frekvensen 0, eftersom ordformen bara förekommer i ett sample. Det är särskilt egennamn som "drabbas" på detta sätt, men naturligtvis även många appellativer, som regel sådana som känns perifera och sällsynta. Man kan konstatera att om *ČS* hade tagit hänsyn till spridningen skulle man säkerligen ha undvikit att t.ex. *kater* erhållit en högre frekvens än *igrat'* (se sid. 819).

Om man bortser från de lägre frekvensområdena kan man med ganska stor säkerhet fastslå att denna lista, varav de 100 första enheterna återfinns i appendix 2, fyller ett högt mått av tillförlitlighet och förhoppningsvis kan förutses förbli i stort sett stabil även på enmiljonnivån. Den lämpar sig väl som "stopplista", alltså som ett redskap t.ex.

vid informationssökning eller textindexering. Vid Slaviska institutionen kommer vi att använda den i pedagogiska syften, nämligen som hjälpmedel vid framställning av halvfabrikat till textordlistor. Liknande frekvenslistor har även framtagits separat för fack- och skönprosa, se appendix 3 resp. 4. Även här är det fråga om endast halva korpusen; någon beräkning av modifierad frekvens på hela korpusen har ännu inte hunnit utföras.

En intressant uppgift som man med lätthet får fram som resultat av en automatisk textbearbetning är förhållandet mellan antalet löpord och antalet olika ordformer i en text. Jag kallar detta förhållande för "token/type-kvot". Kvoten i fråga tycks påverkas av åtminstone följande faktorer:

1. Språktypologi, framför allt med avseende på den morfologiska strukturen. Svenskan tycks sålunda generellt få högre värden än ryskan, vilket också är vad man kan förvänta sig med tanke på att ryskan är mer formrik (har ett större förråd av ordformer) än svenskan. På nivån 100.000 ord uppvisade sålunda rysk facktext i vår korpus en kvot på 3,70, medan en svensk korpus av motsvarande storlek (en vid Centrum för datorlingvistik framtagen samling av artiklar ur *Forskning och Framsteg*) har en kvot på 6,56. På nivån en miljon ord kan man jämföra vår ryska korpus med Alléns tidningskorpus från 1965; kvoterna är 7,53 resp. 9,69. Här måste man dock väga in ytterligare två faktorer, nämligen att kvoten generellt är högre för homogena korpusar samt för fackprosa (jfr nedan). Båda dessa faktorer bör höja kvoten för tidningskorpusen.

2. Textmassans storlek. Kvoten ökar med antalet löpord. I vår korpus är den vid 100 löpord ungefär 1,1, vid 1.000 1,5, vid 5.000 (vår samplestorlek) i genomsnitt 2,06, vid 250.000 4,87, vid en halv miljon 5,76 och vid

en hel miljon, som nämnts, 7,53. De båda sistnämnda kvoterna gäller heterogena textmassor, alltså med såväl fack- som skönprosa; för homogena textmassor blir kvoterna något högre.

3. Enligt Arapov (sid. 31) är en viktig faktor också om vi inom ramen för en given textmassa rör oss med en hel text eller en korpus sammansatt av flera texter: "[...] v celostnom tekste raznoobrazie slov suščestvenno bol'še, čem v korpuse tekstov togo že ob"ema". Att verifiera detta påstående med material från vår korpus är inte så lätt, eftersom den saknar långa sammanhängande texter. Det går i alla fall att konstatera att de sampler som är sammansatta av fler texter inte har färre ordformer per löpord (alltså högre token/type-kvot) än sampler innehållande endast en text.

4. En iakttagelse som i viss mån tycks motsäga Arapovs påstående är att stilistiskt homogena texter har en högre kvot än stilistiskt heterogena. Den nedre resp. övre hälften av vår korpus, vilka alltså var för sig är sammansatta av lika delar fack- och skönprosa, har en genomsnittlig kvot på 5,76 (5,71 resp. 5,80), medan medelvärdet av kvoterna hos den halva miljon som utgörs av enbart fackprosa och den halva miljon som utgörs av enbart skönprosa är 6,3 (6,5 resp. 6,1).

5. Även den funktionella stilen i sig har en viss betydelse i sammanhanget. Dichotomin fack-/skönprosa, som torde vara viktigare än varje annan stilistisk gräns som kan dras i korpusen, har en liten, men ändå påvisbar inverkan på kvoten i fråga. Något överraskande visar det sig att vid en mindre textmassa använder fackprosa fler ordformer än skönprosa, medan det vid en större textmassa förhåller sig tvärtom. På nivån 5.000 har fackprosa en kvot på 1,99 och skönprosa 2,14, men på nivån 500.000 är de inbördes storleksförhållandena på kvoterna omkas-

tade: fackprosa 6,5, skönprosa 6,1. Skönprosan använder sig alltså mer intensivt än fackprosan av de mest frekventa orden; medelvärdet i hela korpusen på de 15 högsta frekvenserna är för fackprosa 5.474, men för skönprosa 6.644, alltså en mycket tydlig skillnad. Å andra sidan har skönprosan s.a.s. en större total arsenal av ordformer att ösa ur än fackprosan; antalet ordformer med frekvensen 1 är sålunda klart större i skönprosan än i fackprosan (48.376 resp. 42.634, även detta räknat på en miljon ord).

6. Eftersom kvoten är ett mått på ordrikedomen i textmassan borde det föreligga stora skillnader mellan individuella stilar, alltså språket hos olika författare. I vår korpus finns författarnamnet inkodat endast för de skönlitterära texterna. Här varierar kvoten från 3,05, som gäller ett barnbokssample av Kaverin, till 1,76, ett sample med text av Tendrjakov. Till författare med hög kvot hör Simonov (2,63), till sådana med låg kvot Nagibin (1,88).

Författare som är representerade med mer än ett sample kan studeras med avseende på spridningen av kvoten. Normalt är spridningen liten, jfr t.ex. följande serie hos Rasputin: 2,29, 2,27, 2,25. Men ett undantag är Granin, som i ett sample når kvoten 1,82 och i ett annat 2,34.

En lexikalisk jämförelse mellan fack- och skönprosa ger vid handen att skillnaderna är betydande. Många ordformer i vår korpus är starkt snedfördelade mellan fack- och skönprosa. Hur det i detta avseende förhåller sig med de mest frekventa orden på halvmiljonnivån framgår av appendix 5, där det görs en direkt jämförelse av den inbördes rangordningen mellan ordformer i fack- resp. skönprosa; för ordformer som har sin partner utanför detta rangområde anges partnerns rang (orden *SSSR* och *menja* hamnar mycket långt ned i rang, vilket markeras med

asterisk).

Snedfördelningen gäller, som synes, inte endast semantiskt laddade ord, främst substantiv, där man själv omedelbart inser t.ex. att *process* och *tetja* frekvensmässigt hör hemma i olika stilar, utan också "strukturord" av typ prepositioner och pronomen. Sålunda är prepositionerna *pri* och *dlja* mycket mer frekventa i fackprosa, medan pronomenformerna *on* och *ona* är vanligare i skönprosa.

Det kan vara av intresse att göra en jämförelse mellan de tio med avseende på modifierad frekvens vanligaste substantiviska ordformerna i fack- resp. skönprosa. Dessa är i fackprosa: *vremja, SSSR, let, raz, goda, delo, žizni, godu, raboty, čeloveka.* I skönprosa når följande former högst frekvens: *raz, vremja, glaza, žizn', den', čelovek, žizni, let, lico, ljudi.* Gemensamma i dessa båda listor är, som synes, ordformerna *vremja, let, raz* och *žizni.* Dessutom är lexemet *čelovek* företrätt på båda ställena, men med olika ordformer.

Eftersom det semantiska innehållet är knutet huvudsakligen till lexemet, ej till ordformen, skulle man förvänta sig att olika ordformer tillhörande samma lexem förhåller sig på samma sätt till dichotomin fack-/skönprosa. Så är inte alltid fallet. Mellan formerna *rek* och *reke* råder sålunda helt omkastade förhållanden: *rek* har den modifierade frekvensen 19.30 i fackprosa och endast 1.59 i skönprosa, medan motsvarande värden för *reke* är 1.43 resp. 18.42. Den "intraparadigmatiska" semantiska kategorin numerus tycks här spela en viss roll. Detta kanske också kan förklara att pronomenformen *oni* inte är snedfördelad, till skillnad från *on* och *ona.*

När man läser frekvenslistor gjorda separat över fack- och skönprosa får man mestadels sin intuitiva uppfattning bekräftad. För att i någon mån undersöka hur väl språkbärarens intuition stämmer överens med korpusen på denna punkt har ett anspråkslöst test genomförts. För detta ändmål utvaldes 90 ordformer med relativt hög totalfrekvens,

varav en tredjedel enligt korpusen är extremt snedfördelade mellan fack- och skönprosa, en tredjedel har moderat och en tredjedel obetydlig snedfördelning. Fyra infödda ryssar fick sedan för varje ordform gissa om den huvudsakligen förekommer i fackprosa eller i skönprosa eller om den är jämnt fördelad mellan dessa båda textkategorier.

Det genomsnittliga antalet rätta gissningar per person blev c:a 55 (mot 30 vid slumpmässig gissning). I de 29 fall då samtliga gissade rätt rörde det sig huvudsakligen – i 24 fall – om extremt snedfördelade ordformer. I 6 fall förekom det att samtliga gissade på samma sätt, men fel (enligt korpusen). Sålunda trodde samtliga att *pered* var jämnt fördelat, medan det i själva verket förekommer oftare i skönprosa. Ordet *drug* ansågs felaktigt som "skönlitterärt", men här förbisågs troligen ordförbindelser som *drug druga*, vilka gör ordet jämnt fördelat. Hos samtliga försökspersoner var det mycket vanligare att man gissade fel mellan jämn och ojämn fördelning än att man tog ett fackprosaord för skönprosaord eller vice versa.

Arbetet med att bearbeta hela korpusen har, som torde framgå av ovanstående framställning, endast inletts. Redan föreligger dock en total konkordans på grafordsnivå (dvs ordformsnivå). Genom ett programmeringsfel har denna konkordans tyvärr blivit något defekt – den kommer om c:a ett år att ersättas av en korrekt version – men den är ändå fullt användbar och torde vara av stort värde för forskare och lärare som behöver excerpera ryska exempel. Nästa steg i projektarbetet är nu att lemmatisera den stora ordformsfilen (132.771 enheter), som utgör ett resultat av bearbetningen av korpusens hela textmassa. Detta kommer förhoppningsvis inom ett år att ge till resultat dels en konkordans på lemmanivå (alltså lexemnivå), dels ett nytt ryskt frekvenslexikon.

Referenser

Allén, Sture *et al.*, *Nusvensk frekvensordbok*, Stockholm 1970.
Arapov, M. V., *Kvantitativnaja lingvistika*, Moskva 1988.
Častotnyj slovar' russkogo jazyka, red. av L. N. Zasorina, Moskva 1977.

Appendix 1

FACKPROSANS FÖRDELNING PÅ TEMAN
(siffrorna till höger anger antal sampler)

1. Biologi	6	14. Medicin/hälsovård	4
2. Data	2	15. Miljö	4
3. Ekonomi	4	16. "Partiliv"	4
4. Energi	4	17. Psykologi	2
5. Fysik	4	18. Rymdforskning	4
6. Försvar	2	19. Sociala frågor	4
7. Geovetenskap	4	20. Sport	4
8. Historia	4	21. Språkvetenskap	2
9. Ideologi/	6	22. Teknik	6
10. Jordbruk	6	23. Utbildn./pedagogik	4
11. Juridik	2	24. Utrikespolitik	6
12. Kemi	4	25. Vardagsliv	4
13. Kultur	4		

SKÖNPROSANS FÖRDELNING PÅ FÖRFATTARE

Fyra författare är representerade med vardera fem sampler:
Granin, D., Kaverin, V., Tendrjakov, V., Trifonov, Ju.

Sex författare är representerade med vardera fyra sampler:
Astaf'ev, V., Ganina, M., Iskander, F., Nagibin, Ju., Rasputin, V., Solouchin, V.

Åtta författare är representerade med vardera tre sampler:
Baklanov, G., Bitov, A., Grekova, I., Gončarov, Ju., Kazakov, Ju., Tokareva, V., Tolstaja, T., Simonov, K.

Tio författare är representerade med vardera två sampler:
Abramov, F., Belov, V., Bondarev, Ju., Vasil'ev, B., Vorob'ev, E., Zalygin, S., Koževnikova, N., Lidin, V., Lichanov, A., Šmelev, N.

Tolv författare är representerade med vardera ett sample:
Ajtmatov, Č., Bek, A., Dubov, N., Gladyšev, N., Zorin, L., Kataev, V., Kočnev, M., Paustovskij, K., Pogodin, R., Pristavkin, A., Troepol'skij, G., Ščerbakova, G.

Appendix 2

DE 100 VANLIGASTE ORDFORMERNA PÅ HALVMILJONNIVÅN
ENLIGT MODIFIERAD FREKVENS
(MF = modifierad frekvens, TF = total frekvens)

	MF	TF		MF	TF
и	18817.18	19277	время	619.14	655
в	15636.84	15941	были	596.64	641
не	8493.20	8829	ли	577.04	622
на	8234.84	8443	во	576.31	605
с	5336.95	5457	где	572.95	610
что	5031.59	5221	быть	572.27	603
а	3745.38	3895	да	556.20	612
как	3166.05	3275	нет	555.40	596
к	2890.32	2970	без	550.88	585
он	2838.28	3132	есть	533.60	568
но	2765.13	2866	раз	516.17	550
по	2651.19	2752	ему	510.58	577
это	2227.35	2310	себя	500.39	539
все	2113.65	2211	них	492.42	523
из	2085.97	2160	этом	492.20	522
его	2030.81	2154	тем	484.87	519
от	1944.20	2012	более	476.84	522
за	1939.08	2025	этого	476.77	506
же	1717.74	1788	которые	476.23	514
у	1643.80	1744	со	475.11	505
я	1604.36	1887	того	472.01	499
было	1496.86	1614	него	467.48	518
для	1464.64	1562	том	465.10	502
о	1460.78	1547	сейчас	458.54	499
так	1440.64	1509	там	451.48	493
она	1408.26	1609	этот	449.90	479
то	1349.06	1427	мне	440.47	520
еще	1316.17	1386	надо	436.91	473
бы	1304.34	1390	теперь	436.38	472
ее	1252.27	1351	лет	427.96	469
только	1179.06	1238	очень	426.29	477
их	1142.49	1203	будет	425.22	462
они	1126.24	1191	тут	420.48	457
уже	1057.05	1113	всего	417.74	446
когда	1024.06	1096	после	415.87	450
до	1001.18	1064	здесь	414.97	450
мы	939.94	1052	нас	411.44	467
или	927.40	992	меня	410.35	490
если	879.05	936	больше	403.90	432
был	854.59	932	потом	401.98	448
чтобы	738.72	781	эти	400.27	426
при	727.43	807	всех	387.45	414
вот	703.35	749	через	385.83	417
ни	694.90	746	тоже	384.03	422
даже	683.92	724	ты	363.32	475
чем	642.07	677	себе	363.21	395
была	634.71	692	лишь	362.73	396
можно	633.46	682	кто	361.86	400
под	620.75	667	потому	361.84	403
может	620.45	657	ведь	356.43	387

Appendix 3

DE 100 VANLIGASTE ORDFORMERNA I FACKPROSA
ENLIGT MODIFIERAD FREKVENS
(MF = modifierad frekvens, TF = total frekvens)

	MF	TF		MF	TF
в	8789.29	8978	быть	294.22	316
и	8536.97	8725	тем	289.67	313
на	3763.43	3917	даже	287.41	310
не	3106.56	3264	всего	284.95	304
с	2385.62	2464	вот	267.28	291
что	2106.97	2200	ли	266.15	297
а	1540.24	1611	них	263.15	287
к	1415.54	1472	лет	260.89	292
по	1340.78	1413	она	255.99	277
как	1277.76	1340	этого	248.63	265
но	1173.89	1236	здесь	244.00	271
это	1117.11	1166	того	243.56	262
из	1080.86	1140	сейчас	242.71	271
для	1066.81	1127	эти	233.22	250
от	914.29	960	без	233.21	254
за	822.99	890	будет	228.25	257
о	810.17	867	нас	221.01	248
же	766.34	812	ни	220.61	241
все	758.18	801	например	219.33	250
его	717.71	764	раз	216.76	239
их	692.35	733	где	215.53	237
то	593.58	633	лишь	210.60	231
при	565.53	628	нет	210.08	234
так	564.45	595	были	206.95	228
мы	563.27	610	других	204.81	223
у	543.21	589	ведь	201.94	222
только	540.83	573	всех	199.85	220
бы	515.49	554	больше	199.63	219
еще	510.10	540	очень	199.16	225
или	509.10	553	этой	197.12	212
они	497.93	532	надо	196.97	219
до	493.39	545	этот	196.91	211
уже	472.58	504	был	192.15	217
он	472.20	521	я	192.10	225
ее	459.42	491	под	191.97	213
если	447.56	486	со	189.03	208
можно	421.29	458	между	186.83	221
более	396.40	424	года	186.53	215
чем	368.64	396	об	185.66	204
было	360.09	386	несколько	182.58	205
чтобы	341.18	367	после	182.45	205
когда	339.80	369	дело	182.16	202
которые	327.44	355	сегодня	181.69	204
время	323.38	348	много	174.02	192
может	320.90	348	однако	173.05	193
во	316.13	334	была	167.72	185
том	312.30	339	жизни	164.79	197
есть	310.24	336	году	163.67	192
СССР	309.29	362	нашей	161.31	182
этом	301.94	321	работы	160.37	179

Appendix 4

DE 100 VANLIGASTE ORDFORMERNA I SKÖNPROSA
ENLIGT MODIFIERAD FREKVENS
(MF = modifierad frekvens, TF = total frekvens)

	MF	TF		MF	TF
и	10172.30	10552	него	367.28	406
в	6834.87	6963	ты	353.03	456
не	5377.66	5565	где	345.85	373
на	4397.93	4526	мы	340.94	442
с	2918.43	2993	себя	340.86	369
что	2877.10	3021	потом	338.49	372
он	2410.51	2611	там	333.19	364
а	2170.04	2284	нет	331.58	362
как	1867.88	1935	без	304.86	331
но	1560.45	1630	тут	302.54	329
к	1441.27	1498	тоже	294.67	323
я	1420.04	1662	ли	292.10	325
все	1347.11	1410	ничего	289.17	314
его	1294.14	1390	теперь	288.66	317
по	1268.33	1339	раз	286.33	311
она	1161.98	1332	может	284.52	309
было	1150.83	1228	время	281.01	307
за	1086.47	1135	со	275.50	297
у	1085.89	1155	быть	265.30	287
это	1076.88	1144	чем	262.32	281
от	1002.33	1052	потому	256.13	291
из	976.50	1020	себе	253.00	276
же	924.75	976	во	249.28	271
так	861.57	914	нее	244.69	280
еще	790.53	846	этот	243.02	268
ее	773.62	860	сказал	242.70	281
бы	764.46	836	ей	235.84	276
то	728.29	794	кто	229.33	256
когда	671.61	727	надо	225.40	254
был	661.11	715	перед	225.36	247
о	617.37	680	вы	225.17	268
только	616.13	665	через	223.26	243
они	605.04	659	тогда	219.90	242
уже	563.58	609	после	219.47	245
до	482.72	519	них	217.32	236
ни	466.13	505	того	217.27	237
была	462.74	507	этого	216.73	241
их	433.38	470	всегда	212.67	233
вот	422.53	458	вдруг	211.86	237
ему	420.84	473	есть	211.46	232
под	420.44	454	очень	208.16	252
если	407.74	450	глаза	204.47	226
да	404.37	447	ну	203.58	228
для	402.35	435	можно	202.23	224
или	392.57	439	сейчас	199.10	228
даже	382.34	414	больше	192.46	213
чтобы	380.59	414	ним	191.06	217
были	378.42	413	жизнь	189.80	217
меня	375.94	444	один	189.40	210
мне	369.77	440	день	188.49	211

Appendix 5

JÄMFÖRELSE MELLAN FACK- OCH SKÖNPROSA AVSEENDE RANGORDNINGEN HOS DE 50 MEST FREKVENTA ORDFORMERNA