Ludmila Ferm och Lennart Lönngren

RYSK TEXTKORPUS. PRESENTATION AV ETT FORSKNINGS-PROJEKT

Inledning

I Uppsala pågår sedan ungefär ett år arbete med att sammanställa en rysk textkorpus som är avsedd att omfatta 1 miljon löpord. Projektet drivs f.n. vid Centrum för datorlingvistik, där bl.a. forskningsingenjörerna Valentina Rosén och Margareta Sjöberg medverkat, men kommer nu att överflyttas till Slaviska institutionen. Vid korpusens uppläggning har det varit nödvändigt att ta ställning till en rad olika frågor, såsom vilken typ av språk korpusen skall representera; hur man skall fördela textmassan över stilar, genrer och ämnen; vilken bakre tidsgräns som skall sättas och om den skall vara samma eller olika för de olika texttyperna. Andra frågor gäller korpusens allmänna struktur, dvs uppdelning i sampler och texter/textutsnitt samt längden på dessa. Vissa av dessa ställningstaganden måste vara definitiva på ett tidigt stadium, medan andra kan revideras under arbetets gång.

Nedan följer en beskrivning av korpusen bl a med avseende på de nämnda aspekterna jämte motiveringar för de beslut som vi hittills har fattat.

Typ av språk. Yttre avgränsning

När det gäller val av språk som skall vara representerat (skriftspråk/talspråk, rikssprak/dialekt, funktionella stilar, avgränsningen i tid), ansluter sig föreliggande korpus ganska nära till den välkända engelska Brown—korpusen. Vi utesluter således såväl poesi som drama. Även beträffande dialoger har vi tagit samma ställning som i Brown–korpusen, dvs dialogrika skönlitterära texter har undvikits. Syftet har varit att få med ett ganska brett spektrum av det moderna ryska språket, dock helt inom ramarna för det s.k. "kodifierade litteraturspråket" ("kodificirovannyj

literaturnyj jazyk"), dvs den ryska som publiceras och läses i Sovjetunionen i dag. Talad ryska har lämnats utanför korpusen, dels därför att det innebär ett komplicerat arbete att samla in sådant material och få det överfört till skriftlig form, dels därför att det är en nackdel om korpusen blir språkligt alltför heterogen. Vi siktar alltså mot en normalspråkskorpus som kan fungera som en standard och bakgrund, mot vilken andra typer av språk kan kontrasteras.

Att åstadkomma en korpus som till hundra procent innehåller "ren" normalprosa är dock svårt. I skönlitterära dialoger, särskilt från lantbruksmiljö, är det mycket vanligt att talspråkliga eller dialektala former finns insprängda. Att stympa sådana texter skulle innebära ett brott mot en annan av våra principer, nämligen att bevara texterna intakta inom ramen för det valda utsnittet.

Ytterligare begränsningar är att vi har uteslutit emigrant– och samizdatlitteratur samt – givetvis – översättningar. Huvudprincipen har varit att alla författare skall ha ryska som modersmål. I enstaka fall kan undantag göras när det gäller allmänt erkända skönlitterära författare som skriver både på ryska och på sitt modersmål, t.ex. Ajtmatov.

Fördelning på stilar, genrer och tematik

Ofta har "allmänna" korpusar innehållit mer fack– än skönprosa, men exempel på motsatsen finns: i Josselson frekvensordbok från 1953 överväger skönlitteratur. Vi har beslutat att vår korpus skall ha lika delar av båda. Motiveringen för denna relativt stora andel skönlitteratur ligger huvudsakligen i att framför allt rysk tidningsprosa starkt skiljer sig från den västliga; den är mindre varierad, mer stereotyp. För att även vardagslivet skall återspeglas tillräckligt i språket behövs en stor andel skönlitteratur.

Fackprosan hämtas i ungefär lika mån från tidskrifter och dagstidningar. Den utgörs genomgående av artiklar av begränsad längd. Monografier och läroböcker utesluts, eftersom det med hänsyn till den snäva tidsgränsen (se nedan) är svårt att täcka större områden med den typen

av texter.

Skönlitteraturen hämtas från böcker samt i något mindre utsträckning från tidskrifter. Den tas från såväl noveller som romaner. Här har vi inte funnit anledning att eftersträva någon bestämd fördelning.

Beträffande tematik lägger vi ned stor möda på att täcka in så många områden som möjligt. Detta underlättas av att samtliga texter och textutsnitt är relativt korta. Vi har för fackprosans vidkommande gjort en systematisk indelning i ämnesområden och viktat dessa mot varandra. Inom skönlitteraturen vill vi givetvis ge större utrymme åt de bästa och mest lästa författarna, utan att dock ensidigt lägga tonvikten vid arbeten som varit politiskt kontroversiella och av denna anledning uppmärksammade. Samtidigt försöker vi spegla hela spektret av genrer och miljöer, bl a barn– och ungdomslitteratur, science fiction, "stadsprosa" med olika stadsmiljöer (fabriker, byggen, forskningsinstitut, skolor osv), landsbygdsprosa, krigsskildringar, naturskildringar etc.

Tidsgräns

Denna fråga hör i viss mån samman med texttypen. När det gäller den stora skiljelinjen, den mellan fack– och skönprosa, bör man nämligen ta hänsyn till att fackprosan, åtminstone lexikaliskt, åldras snabbare än skönlitteratur. Vi har därför valt att ta med skönlitterära texter från 1960 och framåt, medan för fackprosans vidkommande år 1985 satts som bakre gräns. De mest efemära texterna, dagstidningarnas, hämtas från 1987 och framåt.

Dessa tidsintervall gör att den aktuella korpusen kommer att radikalt profilera sig gentemot tidigare existerande ryska korpusar, av vilka ingen innehåller texter som är senare än från 60–talet; Zasorinas korpus, vilken ligger till grund för hennes *Častotnyj slovarʹ russkogo jazyka* (1977), omfattar sålunda texter från början av detta sekel (Lenin, Gorkij) fram till 1968. Särskilt vad gäller fackprosa kommer korpusen att bli såväl språkligt homogen som tematiskt aktuell och diversifierad.

Struktur

Liksom många andra korpusar skall även vår vara uppbyggd av textmoduler (sampler) med fixerad längd, bl.a. för att lättare få fram statistiska undersökningar, t ex vad gäller ordens spridning. En vanlig samplelängd har, åtminstone tidigare, varit 2.000, även om trenden går åt något större sampler. Vi har bestämt oss för 5.000 löpord per sample. Detta medför att andelen obeskurna texter kan hållas högre. För vissa syften, av vilka några blivit aktuella på senare tid, är det en fördel med hela texter. Vid traditionella undersökningar av kvantitativa egenskaper hos hela språket samt hos funktionella stilar, där en naturlig tillämpning är framställning av frekvenslistor och –ordböcker samt över huvud taget framtagning av statistiska medelvärden, spelar textstrukturen föga roll. För textlingvistiska och textspecifika undersökningar, däremot, för bestämning av lexikalisk–semantisk koherens samt ämnes– och författarberoende avvikelser från medelvärden vad gäller både lexikaliska och strukturella egenskaper (t.ex. LIX–värde), är det viktigt att ha en korpus bestående av hela texter.

Med en samplestorlek på 5.000 löpord kommer hela korpusen att bestå av 200 sampler. Ett sample kan bestå av en eller flera texter, i senare fallet måste dock alla texter vara inom samma ämnesområde (gäller fackprosan) eller av samma författare (gäller skönprosan). Endast en text kommer att beskäras inom samma sample. En strävan är att hålla varje sample relativt homogent, dock med hänsyn tagen till att ifrågavarande ämnesområde måste täckas väl.

Representation och inmatning

Principer för textens representation i maskinläsbar form har utarbetats av Valentina Rosén. Textbearbetningsprogrammet framtvingar viss – men ganska obetydlig – redigering av originaltexten. Stanskonventionen utgår från vetenskaplig translitterering, dock med vissa begränsningar: endast ett–till–ett förhållande tillämpas och givetvis får inga diakritiska tecken förekomma: exempelvis använder vi i stället för ja, ju, š och č

respektive ä, ¨, w och å. Inmatningen har hittills huvudsakligen skett för hand. Ansträngningar har gjorts för att få till stånd optisk inläsning av åtminstone en del av materialet. Resultaten är dock hittills otillfredsställande.

Bearbetningar av korpusen

Efter inmatning av materialet kommer korpusen att bearbetas i viss utsträckning. Vi planerar att härvid utnyttja ett programpaket, "Textpack", som har utarbetats vid Centrum för datorlingvistik. Förutom diverse kvantitativa data kommer en grafordskonkordans samt en frekvenslista över graford (en sådan saknas för närvarande vad gäller ryska språket) att framställas.

Möjligheten att göra vidare bearbetningar inom den planerade projekttiden (se nedan) beror bl.a. på i vilken grad inmatningsprocessen kan automatiseras. Följande undersökningar är tänkbara:

1. Kvantitativa jämförelser på grafordsnivå mellan olika delkorpusar, avgränsade enlig kvantitativa eller kvalitativa kriterier. Det kan gälla känsligheten hos vissa kvantitativa parametrar, t.ex. förhållandet mellan totala antalet löpord och antalet olika ordformer, med avseende på total textmassa, vidare framräkning av LIX–värden o. dyl. hos olika funktionella stilar, genrer, (skönlitterära) författare etc.

2. Undersökningar på lemmanivån. Det är först på denna nivå som semantiskt–lexikaliskt relevanta frekvensstudier kan göras. Även här blir jämförelser mellan den totala korpusen och olika slags delkorpusar av intresse. Det ser för närvarande ut som om vi skulle kunna använda samma lemmatiseringsprinciper som hos Zasorina, varför frekvensjämförelser med detta lexikon bör kunna göras (givetvis med beaktande även av ordens spridning).

Vid lemmatiseringsarbetet kommer en rysk ordformsfil, som finns vid Centret och som omfattar ca 60.000 enheter, att utnyttjas. Denna,

som är ett resultat av textbearbetningar med det morfologiska analysprogrammet "Autlex" (Sågvall, *A System for Automatic Inflectional Analysis*, 1973), innehåller kopplingar mellan ordform och lemma.

3. En annan möjlighet är taggning av ordformerna, dvs markering av vissa grammatiska kategorier m.m. Sådan information, som kan vara viktig för bl.a. syntaktiska studier av korpusen, finns redan inlagd i den ovannämnda ryska ordformsfilen. – När det gäller såväl lemmatisering som taggning, vilka innebär stora arbetsuppgifter och därför kanske inte helt kan rymmas inom ramen för projektet, kan förutom ordformsfilen även interaktiva lemmatiseringsprogram och morfologiska analysprogram, främst "Autlex", komma till användning.

4. Markering av ord som inte förekommer i ryska lexika. En sådan bearbetning har gjorts i Zasorinas frekvensordbok. Det gäller här inte bara slang och dialektismer utan i minst lika hög grad specialtermer inom delvis nya och expanderande kunskapsområden.

Projektets fortsättning

Till dags dato (maj 1988) har ca. 200.000 ord inmatats, ungefär jämnt fördelade mellan skön– och fackprosa. Medel har sökts och erhållits (av HSFR) för ytterligare tre års arbete. Eftersom det normalt tar lång tid att lägga upp en textkorpus kan det vara motiverat att sätta upp ett etappmål. I föreliggande fall har vi satt en halv miljon löpord som etappmål. Meningen är att korpusen redan när detta mål är uppnått skall kunna användas för vissa undersökningar; den skall vara mångsidigt sammansatt och ha samma struktur, fördelning på genrer och tematiska områden etc. som den totala korpusen. Etappmålet en halv miljon ord beräknas gott och väl uppnås efter halva treårsperioden, dvs i början av

år 1990.

Under senare hälften av projekttiden kommer vi att arbeta parallellt med dels inmatning upp till 1 miljon ord, dels ytterligare bearbetningar av redan inmatad text.

Det är vår förhoppning att denna korpus kommer att spegla det aktuella språket och utgöra ett rikt material för datorstödd utforskning av modern ryska.