Lennart Lönngren

HALVAUTOMATISK FRAMSTÄLLNING AV TEXTORDLISTOR I RYSKA

Den snabba samhällsutvecklingen i Sovjetunionen har accentuerat behovet av att relativt ofta byta ut framför allt fackbetonade texter i ryskundervisningen. Detta är i och för sig inget större problem om man nöjer sig med att kopiera artiklar ur tidningar och tidskrifter och låter eleverna läsa dessa på traditionellt sätt, dvs genom att slå upp okända ord i lexikon och eventuellt föra glosböcker för att nå bättre inlärningseffekt. Men det förekommer inte sällan dels att texterna används som material för diverse övningar, t.ex. översättningsövningar, dels att de – framför allt på lägre nivåer – förses med rysk-svenska ordlistor. Den sistnämnda typen av hjälpmedel efterfrågas ofta av eleverna, men man möter också (mest bland lärarna) uppfattningen att textordlistor kan vara ett hinder för en "rätt" uppbyggnad av ordförrådet.

Vid Slaviska institutionen i Uppsala har textordlistor använts sedan länge. Ofta har dessa varit framställda på rent subjektiv väg vad gäller urval, grammatisk information etc. Men det har också förekommit försök att framställa ordlistor på ett mer systematiskt sätt (i början av 1970-talet framställdes ett text- och ordlistpaket av Anna Sågvall Hein, vilket fortfarande används; se även Lönngren, L., Rak, J. & Skov-Larsen, J., "Om framställning och användning av ordlistor i tjeckiska", *Slovo* 11, 1975, s. 14-23). En viktig del av dessa strävanden har varit att man har försökt att få stadga i urvalsprocessen genom att ha en basordlista som grund.

I och med att man investerar tid i att framställa ordlistor till texter för undervisning eller självstudium blir det inte lika lätt att förnya detta material. Ett sätt att övervinna denna svårighet kan vara att rationali-

sera själva arbetet. Genom projektet "Rysk textkorpus" (1 miljon l ord, hälften fackbetonad och hälften skönlitterär text, beskrivningar projektet finns i nr 36 och 39 av *Slovo*) har nya möjligheter öppnat den riktningen. Jag har vid institutionen initierat ett pedagogiskt vecklingsprojekt som syftar till att framställa "halvfabrikat" till te ordlistor. Förutom jag själv har Tamara Lönngren medverkat i de projekt.

Förutsättningen är att texterna befinner sig i maskinläsbar for dvs på något sätt är inmatade i en dator. Vi har hittills arbetat m texter som redan finns tillgängliga i den ryska textkorpusen. Det rör om c:a 600, av vilka facktexterna är större till antalet men mindre i omfånget än de skönlitterära texterna. Inget hindrar dock att m skriver in nya texter på samma sätt som tidigare; så småningom kans man även kan utöka volymen och minska kostnaderna genom opti inläsning.

Projektets utgångspunkt är att textordlistor är ett effektivt hjäl medel vid textläsning. Det innebär en tidsvinst att eleven inte behöv slå så ofta i ordböcker. Själva den tid som åtgår till sökandet kan kna past anses väl använd. Dock kan naturligtvis inte en textordlista ge i den information om ordet som finns i ett stort lexikon. Beträffande ordlistor som framställts inom projektets ram har jag tagit det radika beslutet att ge minimal information: i den vänstra kolumnen enda ordet i dess lexikonform helt utan grammatisk information och i d högra endast den svenska ekvivalent som är aktuell i texten. Huvudsy tet är att man lätt skall ta sig vidare i texten utan att fastna.

Den bakomliggande filosofin kan formuleras i två punkter: a) det i bättre att i början av studierna läsa en stor mängd lätt text än en lit mängd svår text; b) det är bättre att ägna mer tid åt högfrekventa än i lågfrekventa ord. Dessa punkter hänger i någon mån samman. D dominerande syftet med textkurser i språk måste vara inte att tillägn

sig innehållet i de lästa texterna utan att öka beredskapen att läsa nya texter. Och det bästa sättet att utöka ordförrådet och befästa ordens form och innehåll är att möta dem i ständigt nya kontexter.

För att textordlistorna skall tjäna sitt syfte måste urvalet av ord så nära som möjligt svara mot läsarens informationsbehov, dvs de skall helst innehålla alla okända ord och endast dessa. För att uppnå denna effekt kan man ur den fullständiga listan över textens ord rensa bort alla högfrekventa. Om detta arbete skall göras automatiskt måste det i huvudsak utföras på ordformsnivå; i varje fall har det varit den enda tekniskt framkomliga vägen i detta fall.

Man behöver alltså, enkelt uttryckt, jämföra listan över den aktuella textens alla ordformer med en lista över de mest frekventa ordformerna i ryskan. I projektet "Rysk textkorpus" har vi f.n. tillgång till dels en lista, baserad på totalfrekvens, som täcker hela korpusen, dels en lista, baserad på modifierad frekvens, vilken täcker halva korpusen (men fortfarande med samma fördelning: hälften fack-, hälften skönprosa). Vid framtagningen av denna lista har vi fått hjälp av forskningsingenjör Bengt Dahlqvist på Centrum för datorlingvistik i Uppsala (UCDL). Den modifierade frekvensen tar hänsyn till såväl totalfrekvens som spridning över samplerna (en sample = en enhet på 5.000 löpord i korpusen). Vi har stannat för den senare listan såsom varande den mest tillförlitliga och vi har kapat av den vid (mer exakt: strax nedanför) gränsen 10.000; detta innebär att inget ord i denna lista har en lägre totalfrekvens än 6. Textordlistan jämförs med frekvensordlistan, och alla ord i textordlistan som även finns i frekvensordlistan rensas bort.

Ett annat krav på textordlistan är att sökandet efter ord elimineras till ett minimum. Detta sker genom att orden i textordlistan ges i samma följd som de förekommer i texten (avser första förekomststället). Detta kan ordnas automatiskt genom att textbearbetningsprogrammet noterar löpordsnummer för första förekomst av varje ordform. Därefter

kan orden sorteras efter denna siffra (i stigande följd).

Jag skall nu övergå till att beskriva proceduren i detalj. De rutin som har utarbetats omfattar inte endast framtagning av ordlistor ut också förnyad utskrift av själva texterna. Detta kan vara befogat i fall då originalet p.g.a. bristande tryckkvalitet är svårt att kopiera ett tydligt sätt; dessutom ger en laserutskrift större möjlighet att redig ra texten efter pedagogiska behov, t.ex. att accentuera den.

När texten skrivs in sker detta inte strikt efter det tryckta origin let. Om man vill att programmet skall göra rätta meningsavgränsning måste man t.ex. skilja på vanlig meningsslutspunkt och förkortning punkt. Vidare måste gräns mellan textstycken markeras med särski tecken. Programmet reducerar också alla stora bokstäver till sm såvida man inte förhindrar detta (framför allt vid egennamn) genom e särskilt tecken.

Texten skrivs in med ryska bokstäver i ordbehandlingsprogramm T^3. Därefter förs den över till UDACs stordator (operativsystem GUTS) och bearbetas med hjälp av programpaketet TEXTPAC (framtaget vid UCDL av Valentina Rosén och Margareta Sjöberg). E av de resultat man härvid kan få är en lista över alla olika ordformer texten, samt markering av beläggställen, dvs ordformernas ordning nummer i den löpande texten. Vi börjar med att avlägsna alla beläg ställen utom det första samt sorterar listan på ett sätt som stämm överens med vår frekvenslista (TEXTPACK ger det ryska alfabete sorteringsföljd). Därefter kör vi ett program, BFREK (skrivet av Agn Kilar, tidigare forskningsingenjör på UCDL), som avlägsnar alla ordfo mer ur textordlistan, vilka även förekommer i frekvensordlistan. genomsnitt avlägsnas på detta sätt 61 % av orden, men det är gansk stor variation mellan texterna.

Nu vidtar arbetet med att på ett automatiskt sätt i så stor utsträck ning som möjligt återföra ordformerna till deras grundformer. Dett

görs genom en rad kommandoprocedurer (SUBST, ADJ, VERB, ALLM), av vilka de tre första tar hand om var sin ordklass och den sista om vissa allmänna stavningsregler. Dessa procedurer innehåller en mängd kommandon av typen *skogo* → *skij* (dvs en relativt entydig teckensträng i slutet på en böjd form ersätts med motsvarande sträng i ordets grundform). Eftersom det härvid genereras nya former av vilka en del inte finns i den aktuella texten kör vi programmet BFREK ytterligare en gång. Som ett resultat av denna körning brukar ytterligare 3 % av orden avlägsnas.

Den starkt reducerade ordfilen omsorteras nu efter den kolumn där det första beläggstället anges. Orden kommer därvid i den ordning som de förekommer i texten. Därefter förs ordfilen tillbaka till PC–miljö, nämligen till T^3 (via DOS). Här sker en manuell finputsning. De böjda ordformer som fortfarande finns kvar återförs till grundformen. Alla ord, framför allt internationalismer, som skulle få ungefär samma utseende i svensk översättning, avlägsnas. Vidare avlägsnas de flesta egennamn. Förkortningar behöver dock ofta dechiffreras.

Genom denna manuella reducering minskar antalet ord till i genomsnitt drygt 20 % i jämförelse med den ursprungliga ordfilen. Nu återstår endast arbetet att ge svenska ekvivalenter till dessa ord. Denna fas kan ibland inbegripa vissa justeringar även av den ryska kolumnen. Man kan t.ex. ibland vilja ge en ordförbindelse i stället för ett enstaka ord.

Till sist överförs även texten till PC-miljö, sedan den dessförinnan rensats från de ovannämnda speciella inkodningstecknen (förkortningspunkt blir åter vanlig punkt etc). I ordbehandlingsprogrammet sker automatisk styckeuppdelning. Ev. accentuering görs tills vidare direkt på pappersoriginalet.

Hittills har 24 texter med tillhörande ordlistor framställts på detta sätt (två av texterna är nyinmatade). Ovan beskrivna procedurer får anses ganska väl utprovade. Förutsättningar finns därför att vid behov

fortsätta denna enligt min mening rationella produktion av ifrågavara de undervisningsmaterial.

I förlängningen kanske vi också kommer att utnyttja ytterliga fördelar som följer av att man har texten i maskinläsbar form. Man k t.ex. ta fram ett läsbarhetsindex (se Scandinavian PC Systems progra KIX) eller en lista på de mest frekventa ordformerna eller redige texten, t.ex. adaptera den, förse den med luckor för lucktest etc. I tilltalande tanke skulle också vara att få i gång ett samarbete m andra institutioner, speciellt vad gäller den tids- och resurskrävan textinmatningen.