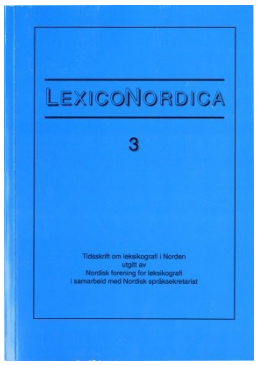
LexicoNordica

Titel: Finlandssvenska, finsk-svensk tvåspråkslexikografi och korpusar

Forfatter: Nina Martola

Kilde: LexicoNordica 3, 1996, s. 105-120

URL: http://ojs.statsbiblioteket.dk/index.php/lexn/issue/archive



Søgbarhed

Artiklerne i de ældre LexicoNordica (1-16) er skannet og OCR-behandlet. OCR står for ’optical character recognition’ og kan ved tegngenkendelse konvertere et billede til tekst. Dermed kan man søge i teksten. Imidlertid kan der opstå fejl i tegngenkendelsen, og når man søger på fx navne, skal man være forberedt på at søgningen ikke er 100 % pålidelig.

Nina Martola

Finlandssvenska, finsk-svensk tvåspråks-lexikografi och korpusar

There is a need for a reasonably large corpus as a tool for the research into the Swedish language as it is used in Finland and as a basis of dictionary compiling. Actually, a corpus of the Swedish used in Finland already exists. It was created by the Department of Scandinavian Languages and Literature of the University of Helsinki and it consists of different types of texts from the early nineties, in other words quite modern language. However, the corpus is far too small, only about two and a half million words, which means that it needs to be enlarged. Such a project will be carried out by the Swedish Department of the Research Institute for the Languages of Finland, if possible in collaboration with the Department of Scandinavian languages and literature of the University of Helsinki. Unfortunately the resources for the project are quite limited, but as it is possible to use partly automatic tagging, an aim of twenty million words is not unrealistic. As for the types of texts, the outlines for the European PAROLE-project will be followed. Other text types will be included in a later phase.

Another corpus project to be carried out is the creation of a parallel corpus with the same texts in Swedish (preferably Standard Swedish) and Finnish. It should consist of Finnish texts with translations into Swedish, Swedish texts with translations into Finnish and, finally, translations from other languages into both Finnish and Swedish. This corpus will be used for finding Swedish correspondences of certain types of Finnish words and constructions that are problematic for the lexicographer.

Bakgrund

Under åren 1990–95 utarbetades manuskriptet till en ny finsk-svensk storordbok som ett samarbetsprojekt mellan Forskningscentralen för de inhemska språken och förlaget WSOY. Boken utkommer enligt planerna hösten 1996. När arbetet inleddes hade redaktionen inte tillgång till en enda korpus. I slutskedet kunde tre olika korpussamlingar utnyttjas; Språkbanken vid Göteborgs universitet och de finska och de finlandssvenska korpusarna vid Helsingfors universitet. En avsevärd utveckling på tre år, alltså. Men trots det kan man knappast kalla arbetet med finsk-svensk ordbok **korpusbaserad lexikografi**, snarare då **korpusstödd lexikografi**.

De närmaste åren kommer forskningscentralen att satsa mycket på korpusarbete och i det följande presenteras denna satsning. En betydligt

mera omfattande finsk korpus än den nu existerande kommer att sammanställas i samarbete mellan Helsingfors universitet och Forskningscentralen för de inhemska språken.

På svenska avdelningen vid Forskningscentralen för de inhemska språken kommer det att satsas dels på en utvidgad finlandssvensk korpus, dels på en parallellkorpus med samma texter på finska och svenska. Det är de två senare projekten som i första hand kommer att behandlas i denna artikel. Det finska korpusprojektet presenteras dock i korthet, eftersom tanken är att samma program och samma kodnings- och söksystem skall utnyttjas i arbetet med den finlandssvenska korpusen.

En nufinsk korpus

För det finska korpusprojektet har projektmedel redan beviljats, och planerna är längre framskridna än för den utvidgade korpusen över finlandssvenska. Forskningscentralen har inlett samarbete med Institutionen för allmän språkvetenskap vid Helsingfors universitet och man kommer att delta i det europeiska PAROLE-projektet. Den nya finska korpusen kommer i första fasen att bestå av 20 miljoner ord och att den kodas i enlighet med TEI-normen enligt samma kriterier som i de övriga deltagarländerna. PAROLE-korpusarna skall enligt överenskommelse bestå av (Le Parole. Technical Annex 1995:25)

58–72% tidningstext
16–22% boktext
4–10% tidskriftsmaterial
8–12% diverse

Som bok räknas allt som har ett ISBN-nummer. Åtminstone i den finska korpusen kommer dock skönlitterär text och icke-fiktion att hållas isär. Småningom kommer korpusen förmodligen att utvidgas till 30–50 miljoner ord. Den skall sammanställas av hela texter av så varierande slag som möjligt. För att ge en heltäckande bild av språket borde korpusen också innehålla texter som inte har underkastats språkgranskning. I något skede vore det viktigt att få med även talat språk.

Den finska korpusen skall kunna utnyttjas för flera olika syften, men de mest omfattande textmassorna behövs för uppdateringen av Suomen kielen perussanakirja ('Finsk basordbok') och för andra nuspråkliga lexikografiska projekt. Korpusarna kommer också att utnyttjas bl.a. av

språkvården, det nufinska grammatikprojektet och av textforskare. (Lehtinen et. al. 1995:29ff.)

Materialet måste antingen kodas morfologiskt i så hög grad som resurserna medger eller också måste man ha tillgång till sökprogram som tillåter sökning på ordstam – detta för att kunna hantera finskans rika böjningssystem (Lehtinen et. al. 1995:69 f.)

Finlandssvenska

Det existerar såsom nämndes i inledningen redan en finlandssvensk korpussamling, men i sitt nuvarande omfång är den alldeles för liten för att ge tillfredsställande sökresultat. Så här ser det ut:

FINLAND SWEDISH CORPUS (FISC)
Department of Scandinavian Languages and Literature, University of Helsinki

# for	Extent (tokens)	Description
1. Daily newspapers (from 1991)	895.000	Hufvudstadsbladet (20 iss.) Vasabladet (7 issues)
2. Literary prose (from 1990–92)	665.000	12 novels or collections of short stories.
3. Non-fiction (from 1990–93)	455.000	Various texts of ordinary factual prose
4. (Texts by) Authorities (from 1990–94)	440.000	Legal texts from the Statute Book of Finland. Various texts produced by public services etc.
5. Spoken language	80.000	Transcriptions of speech
6. Monitor	undefined	Miscellaneous texts

Grupp 5 innehåller utskrifter av sex radiointervjuer och fyra radiodiskussioner. I grupp 6 ingår bl.a. ytterligare tidningstext; Hufvudstadsbladet ca 120.000 ord.

Ingen av de enskilda korpusarna innehåller alltså ens en miljon löpord, vilket betyder att man får alldeles för få belägg också på relativt vanliga ord och uttryck för att kunna dra några egentliga slutsatser. Som ett exempel kan nämnas att en sökning på uttrycket *av misstag* i Språkdatas DN-konkordans på 4,1 miljoner löpord gav 33 belägg medan en sökning på *av misstag* och *i misstag* i det finlandssvenska pressmaterialet gav sammanlagt sju belägg; fem där prepositionen var *av* och två där den var *i*. (Uttrycket *i misstag* är en s.k. finlandism.)

En finlandssvensk baskorpus

Den finlandssvenska korpusen måste alltså utvidgas för att ge mera tillfredsställande sökresultat. Tanken är att Svenska avdelningen vid Forskningscentralen för de inhemska språken skall försöka skaffa projektpengar och att ett samarbete skall inledas med Institutionen för Nordiska språk vid Helsingfors universitet, som hittills har upprätthållit korpusen.

Det första steget blir att helt enkelt bygga ut den existerande korpusen genom att fylla på med mera texter i varje typ, d.v.s.:

1) Mer pressmaterial; dels fler nummer av de nu ingående tidningarna, dels fler olika regionala tidningar. Att pressmaterialet utgör en stor andel är befogat. Dels utgör ju tidningar en mycket stor andel av de texter som trycks per år i ett land, dels hör tidningstexter till de texter som läses allra mest.

2) Fler romaner

3) Mer faktatext av olika slag

Det första steget blir att sammanställa en traditionell korpus enligt samma principer som den finska systerkorpusen, alltså enligt PAROLE-projektet (se ovan). Den existerande korpusen är i ganska hög grad uppbyggd enligt samma principer som Språkbanken i Göteborg, som ju också kommer att ingå i PAROLE-projektet. Om korpusen skall följa samma principer som den finska bör den TEI-kodas.

Förutom tidningstext är broschyrer från olika statliga och kommunala verk en texttyp som produceras på svenska för finlandssvenskar och som läses rätt allmänt. Problemet med dessa är att de för det mesta är översatta. Endast i kommuner med svenska som majoritetsspråk produceras sådant material primärt på svenska. Skillnaderna på de texter som hade producerats direkt på svenska och de som hade översatts från finska i en och samma kommun visade sig, enligt uppgift från forskaren Eivor Sommardahl, vara stora i ett kommunalt klarspråksprojekt som har genomförts i samarbete mellan svenska avdelningen vid Forskningscentralen för de inhemska språken och Finlands kommunförbund. Det är ju i och för sig ingen överraskning. Mera om problemet med översatta texter längre fram.

Det vore naturligtvis önskvärt att också den finlandssvenska korpusen kunde utvidgas till 20 miljoner ord, men det är långt ifrån säkert att det lyckas inom en överskådlig framtid, eftersom det knappast kommer att kunna avsättas lika mycket resurser för den svenska

korpusen som för den finska. Med de redskap som har utvecklats på Institutionen för allmän språkvetenskap vid Helsingfors universitet finns det dock möjlighet till relativt långt driven automatisk kodning, vilket gör att det ändå kan finnas rimliga chanser att också med små personresurser bygga upp en korpus av hyggligt omfång.

Också den finlandssvenska korpusen skall kunna användas för olika syften. För vissa ändamål, t.ex. för syntaktiska studier, borde ett lämpligt urval av texter kodas mer detaljerat, men för de lexikografiska behoven (och i hög grad också för språkvårdens behov) är det bättre att få med en så stor mängd text som möjligt och nöja sig med en grov kodning. Det viktigaste är oftast att helt enkelt få belägg på att ett ord eller uttryck existerar, eller att få ett tillräckligt antal belägg för att det skall gå att dra slutsatser om betydelse(nyanser). (Jfr Sinclair 1992:4).

En kompletterande korpus

Texter producerade av skolelever

Det har skrivits spaltmetervis om finlandssvenskan, men trots det stötte vi i arbetet med den finsk-svenska ordboken ständigt på problemet att finlandssvenskan är synnerligen otillräckligt kartlagd vad gäller regional variation. Det finns en stor excerptsamling över s.k. finlandismer, d.v.s. ord och uttryck som är okända (eller ibland bara föråldrade) i rikssvenskan. Men i regel har de påträffade språkdragen kontrasterats enbart mot rikssvenskan. Någon jämförelse mellan olika regioner i Finland har inte gjorts, och mycket av det som anförs som finlandismer kan vara typiskt för någon mindre region och helt okända i andra regioner. Ju ledigare och mera talspråksnära språket är desto större är naturligtvis variationen. Det skulle behövas "mera kött på benen" därvidlag och för det ändamålet vore en korpus med regionalt förankrade texter ett intressant experiment.

Hur kunde en sådan korpus tänkas se ut? En språkligt relativt "oförstörd" grupp i det här hänseendet, och en grupp som ändå skriver en hel del, är grundskolelever. Ett sätt att komma åt regionalismer vore att utnyttja datoriseringen i skolorna och be svensklärarna (eller varför inte andra intresserade lärare) i ett antal skolor fördelade på olika regioner sända in orättade elevtexter på diskett. Eleverna måste självklart bes om lov och de skall naturligtvis få vara anonyma. Texterna bör kodas dels med avseende på skribentens hemort, dels med avseende på språklig bakgrund – standardspråk, dialekt samt grad av

tvåspråkighet (t.ex. ena föräldern finsk, båda finska, finska kamrater eller släktingar).

Enligt uppgifter från lärarhåll är det stor skillnad på grundskole- och gymnasienivå. Gymnasieeleverna har en tydlig strävan mot och en relativt stark medvetenhet om den skriftspråkliga normen, medan grundskoleleverna skriver ett talspråksnära språk, vilket betyder att de regionala variationerna borde komma bra fram i deras uppsatser. Tar man med uppsatser från bägge nivåerna får man ett gott jämförelse-material och kanske därmed en uppfattning om etableringsgrad för olika regionalismer.

Problemet med ett projekt av det här slaget är att materialet är ytterst arbetsamt att samla in och att förenhetliga. Om man räknar med att målet i första fasen vore en korpus på ca 2 miljoner ord innebär det att man måste bearbeta ca 4.000 elevuppsatser à ca 500 ord. Räknar man med nio regioner skulle det betyda 445 uppsatser per region. Fördelar man det på flera skolor är det i och för sig kanske inte en omöjlighet att få in ett sådant antal texter i elektronisk form inom rimlig tid. Men det betyder med andra ord att varje region representeras av 222.500 löpord, vilket är alldeles för litet. För att se om ett sådant här företag över huvud taget är möjligt och försöka beräkna vilka resurser som krävs är det säkert skäl att först genomföra ett mindre pilotprojekt.

Talspråk

Som framgick av översikten ovan ingår det i den finlandssvenska korpussamlingen en korpus bestående av 80.000 ord transkriberat tal. Radiospråket representar dock en relativt formell nivå, och för att undersöka regionala skillnader borde man få till stånd en korpus också med informellt tal. Tillsammans med elevuppsatskorpusen kunde en sådan talspråkskorpus ge en bättre grund för beskrivningen av informell finlandssvenska och regionala skillnader. Att sammanställa en talspråkskorpus är dock tidskrävande, så det kommer att bli en senare uppgift.

Texter riktade till barn och ungdom

Intressanta texter i framtidsperspektiv är sådana texter som riktar sig till barn i skolåldern. Det språket spelar säkert en inte så obetydlig roll när det gäller att forma det finlandssvenska "standardspråket" som det kommer att se ut när dagens skolelever vuxit upp – citationstecken

eftersom någonting sådant som ett finlandssvenskt standardspråk egentligen inte existerar. Främst skulle skolböcker kunna komma i fråga; det produceras en hel del läromedel på svenska i Finland, även om mycket också är översatt från finska.

Det skrivs också en del barn- och ungdomsböcker på svenska i Finland. Vad gäller konsumerad litteratur spelar de en relativt liten roll, eftersom det sverigesvenska utbudet är många gånger större, men rent språkligt är det ett intressant material. Många barn- och ungdomsförfattare strävar efter att skriva ett målgruppsanpassat språk, så att läsarna skall uppleva språket som sitt. Det här betyder att man antagligen kan träffa på en hel språkdrag som författarna uppfattar som typiskt finlandssvenska i texterna. Här får man dock vara noggrann med urvalet, om det är just finlandssvenska särdrag man vill komma åt. En del författare har i stället en strävan att anpassa sitt språk efter sverigesvenska för att "bibringa barnen ett bra språk".

Översatta texter

Jag har redan varit inne på problemet med att så mycket är översatt i Finland. Översatta texter skiljer sig ju språkligt från orginaltexter på ett språk, och man kan därför argumentera att översättningar från finska till svenska inte hör hemma i en finlandssvensk korpus, att de egentligen inte speglar finlandssvenskan. Å andra sidan vet vi att språkdrag kan sprida sig från språk A till språk B via frekvent översättning och småningom börja uppträda också i texter skrivna direkt på språk B (Gellerstam 1986:92f.). Översatta texter – av bättre eller sämre kvalitet – utgör en avsevärd del av de svenska texter man som finlandssvensk konfronteras med. De kan därför bidra till att språkdrag som är frekventa i just översatta texter småningom befästs i finlandsvenskan. Att översatta texter och icke-översatta texter måste hållas isär i korpusar säger sig självt.

Användningen av den finlandssvenska korpusen

Den finlandssvenska korpusen är tänkt att tjäna flera syften. Inom forskningscentralen kommer språkvårdarna att bli en mycket viktig användargrupp, men korpusen skall också kunna utnyttjas för forskning i finlandssvenska. Dessutom skall den utnyttjas av två lexikografiska projekt.

Finlandssvenskan i finsk-svensk ordbok

Korpusen kommer att vara en tillgång vid framtida uppdateringar och bearbetningar av den finsk-svenska storordboken. Under arbetet med det manuskript som nu är under tryckning var det ett ofta återkommande problem att det inte fanns tillräckligt med finlandssvenska texter i maskinläsbar form för att man skulle kunna räkna med att avsaknaden av belägg faktiskt indikerade att det sökta ordet inte användes. När det gäller den finlandssvenska som används i mera officiella sammanhang går det ibland att hitta belägg på olika ord och uttryck i andra källor. Men långt ifrån alla källor har ju index över viktiga ord och termer, och dessutom är ju inte alla ord man söker terminologiska. Att försöka plöja igenom texter på måfå för att leta efter belägg på något visst ord är ju som att leta efter en nål i en höstack.

Ger man ut en finsk-svensk ordbok i Finland är det alldeles klart att finländska särdrag skall tas med. Sådana finns på alla språkliga nivåer och det är besvärligt att dra gränser för hur de skall redovisas. Nedan några exempel, först ett par officiella termartade benämningar (förkortningarna *Suom*. och *Ruots*. med versaler för 'i Finland' resp. 'i Sverige'), sedan ett kulturbundet fenomen och sist ett talspråkligt uttryck (förkortningarna *suomr*. och *ruotsr*. för 'finlandssvenska' resp. 'rikssvenska').

kantopiiri **Suom.** utbärningsdistrikt -et =; **Ruots.** utdelningsområde -t -n **myös suomr**. **yleisk**.
diplomi-insinööri **Suom.** diplomingenjör -en -er; **Ruots.** civilingenjör
talko|ot[17] **suomr.** talko -t -n (-n -r); **talkootyö** talkoarbete -t -n; ruotsr. arbetsgille -t -n; grannhjälp -en; *järjestimme ~ varaston siivoamiseksi* vi samlade ihop frivilliga för att få lagret städat; vi hjälptes åt för att få lagret städat; **suomr**. vi ordnade talko för att få lagret städat; *oja kaivettiin -illa* vi hjälptes åt att gräva diket; diket grävdes med gemensamma krafter; **suomr**. diket grävdes på talko (som talkoarbete)
simaht|aa[53*F] **väsähtää** slockna[1]; **mennä rikki: vars. ruotsr.** paja1; **suomr**. **laitteista** säga[4] upp kontraktet; *hän -i kesken matkan* han slocknade på halva vägen; *-anut juhlija* en festare som slocknat

Klart är att termartade uttryck som förekommer i officiellt språkbruk skall redovisas. Att också sådana ord och uttryck som används över hela det finlandssvenska språkområdet skall redovisas rådde det i princip enighet om inom redaktionen. Däremot var redaktionen inte

alltid enig beträffande vad som är "allfinländskt". Ju mer man rör sig i riktning mot informellt språkbruk, desto större blir de regionala skillnaderna och desto svårare blir det också att hitta belägg på de ord man söker. Här skulle alltså korpusen med skolelevstexterna och talspråkskorpusen, kunna ge indikationer på hur allmän spridning ord och uttryck har, detta förutsatt att de kan byggas ut tillräckligt.

Ett stort lexikografiskt problem kvarstår dock, oavsett alla korpusar, och det är vad man skall göra i de fall när det inte existerar ett gemensamt talspråkligt ord på finlandssvenska på samma sätt som det många gånger gör i finskan och i rikssvenskan. Skall man redovisa olika alternativ med regionala markeringar av något slag eller skall man gå in för en viss regions talspråk, t.ex. Helsingforsregionens, eller skall man helt och hållet utesluta de finländska varianterna i sådana fall. Ingen av lösningarna är särskilt tillfredsställande.

En ordbok över finlandssvenskan

För några år sedan började en ordbok över finlandssvenskan utarbetas på svenska avdelningen vid Forskningscentralen för de inhemska språken. Det är egentligen inte en ren ordbok, utan det kommer att vara en handbok bestående dels av en textdel med resonerande text, dels en ordboksdel med ca 2.000 uppslagsord. Orden utgörs dels av särfinlandssvenska ord, dels av allmänsvenska ord som helt eller delvis har fått annan betydelse eller användning i finlandssvenskan. Ordboksartikeln i övrigt innehåller antingen en förklaring eller en uppgift om den allmänsvenska motsvarigheten och eventuella uppgifter om bakgrunden till ordet eller användningen. Så här ser det ut i en preliminär version:

Fig. 1. Ur en preliminär version av den finlandssvenska ordboken

Uppslagsorden är hämtade ur ett omfattande register över s.k. finlandismer, d.v.s. ord och uttryck som antingen används bara i finlandssvenskan eller som används i en annan betydelse i finlandssvenskan än i sverigesvenskan. Registret omfattar för närvarande ca 5000 poster och det har tillkommit dels genom att litteratur *om* finlandssvenska har excerperats, dels genom att finlandssvenskt textmaterial excerperats. Excerperingen har gjorts av många olika personer och det excerperade materialet har hållit varierande standard, så det finns upptaget en hel del sådant som verkar vara ytterst tillfälliga bildningar eller direkta språkfel. I allmänhet finns heller ingenting noterat om frekvens eller om regional utbredning. Det vore alltså en stor fördel att ha en tillräckligt omfattande korpus att systematiskt kontrollera registret mot. Samtidigt skulle säkert en sådan genomgång tillföra registret en hel del nya poster.

Registret kunde sedan utgöra utgångspunkten när lemmalistan till den finlandssvenska ordboken utarbetas. Tilltänkta uppslagsordskandidater kontrolleras mot korpusen och ord och användningar av ord som finns väl belagda tas med. På så vis får man en adekvat kärnuppsättning uppslagsord.

Å andra sidan råkar man naturligtvis alltid ut för att det finns bara något enstaka belägg i en korpus eller inga alls, hur stor korpusen än är. De fallen måste man bedöma utifrån andra premisser. En korpus kan knappast vara det allena saliggörande vid ordboksarbete. Den är ju alltid bara ett litet urval av språket, hur omfattande den än är.

För närvarande ligger redigeringsarbetet på is, eftersom inga personella resurser kan reserveras för det. Sannolikt kommer dock den första versionen av den finlandssvenska ordboken att utarbetas utgående från det preliminära manus som nu föreligger, men för framtida upplagor av ordboken kommer korpusen att bli en viktig bas.

Korpusanvändning inom tvåspråkslexikografin

Som tvåspråkslexikograf ställer man sig ibland litet undrande inför debatten om huruvida bara autentiska exempel får användas i en ordbok eller om även redaktionella exempel kan tillåtas. Det är en debatt som bara kan gälla enspråkslexikografin. Tvåspråkslexikografen har inget val. Hur enorma korpusar man än får till sitt förfogande kan man aldrig räkna med att för varje källspråksexempel man väljer ut hitta ett målspråksexempel som är en exakt motsvarighet och som oredigerat kan fungera som översättning. Man blir helt enkelt tvungen att bearbeta exemplen redaktionellt, ibland både käll- och målspråksexemplet, ibland åtminstone endera. Avskalade exempel fungerar dessutom oftast bättre i en tvåspråkig ordbok än obearbetade. Dels ger de generellare information, dels ger de inte upphov till onödiga översättningsproblem som är ovidkommande för det språkliga fenomen man vill belysa och som ibland rent av kan skymma det man ville klargöra.

Korpusarna för käll- och målspråk utnyttjas för olika ändamål. Om man börjar från noll när man utarbetar en ny tvåspråkig ordbok används källspråkskorpusen på i stort sett samma sätt som när man utarbetar en ny enspråkig ordbok. Det här är väl dock relativt ovanligt. Det vanligaste är väl att man utnyttjar ett existerande material, antingen så att en existerande ordbok revideras eller så att man utnyttjar en enspråkig ordbok eller en ordbok till ett annat målspråk som bas. Källspråkskorpusen använder man då i första hand för kompletterande uppgifter; för att få belägg på ordbetydelser eller användningar som man upptäcker fattas eller för att få fler exempel.

Målspråkskorpusen letar man i stor utsträckning i på källspråkets villkor, d.v.s. man letar efter ord, uttryck och konstruktioner som i så hög grad som möjligt motsvarar ett källspråksexempel, och man letar efter belägg på att ett ord som man tänker sig som ekvivalent faktiskt existerar.

En gemensam bas för olika tvåspråkiga ordböcker

En tanke som nu har framkastats på forskningscentralen är att utarbeta en gemensam bas för tvåspråkiga storordböcker från finska till andra språk. Andra målspråk än svenska kommer att bli aktuella framöver, och då vore det viktigt att ta vara på erfarenheterna från det finsk-svenska ordboksprojektet.

Det går naturligtvis inte utan vidare att överta en enspråkig ordbok och "översätta" materialet i den. På det viset uppstår inte en adekvat tvåspråkig ordbok. En tvåspråkig ordbok från språk A till språk B är inte nödvändigtvis heller en bra bas för en ordbok från språk A till språk C, eftersom ett annat målspråk aktualiserar delvis andra kontrastiva aspekter. En gedigen bas kunde däremot åstadkommas genom en syntes av en enspråkig och en tvåspråkig. För varje ny tvåspråkig ordbok som görs kan materialet kompletteras med nya kontrastiva aspekter. På det här viset skulle man upprätthålla en bas som är betydligt mer omfattande än vad de färdiga ordböckerna är tänkta att bli, och som utnyttjas till de delar som det aktuella målspråket förutsätter.

Det första steget skulle alltså vara en syntes av Suomen kielen perussanakirja och Finsk-svensk ordbok. Det senare materialet borde analyseras systematiskt utifrån kontrastiva aspekter, för i sitt nuvarande skick är det behäftat med vissa brister som beror på att en enspråkig ordbok legat till grund. Även om gallringar och tillägg har gjorts under redigeringsarbetets gång är nog materialet i vissa avseenden för "enspråkigt". En jämförelse med andra tvåspråksordböcker skulle säkert ge en hel del matnyttigt i all synnerhet vad gäller exempelmaterialet.

Såsom tidigare nämndes kommer Suomen kielen perussanakirja att uppdateras. Kompletteringarna skall göras utgående från den nya nufinska korpusen. De uppdateringar och kompletteringar som görs för den enspråkiga ordboken bör naturligtvis utnyttjas också för den tvåspråkiga basen, men dessutom bör nog en separat genomgång av korpusarna göras för denna, eftersom bl.a. exempelvalet delvis utfaller annorlunda i så fall, än om man letar efter exempel för en enspråkig ordbok.

För att åstadkomma bra produktionsordböcker vore antagligen en "baklängesgenomgång" av korpusmaterialet givande, d.v.s. att utgående från typiska uttryck på målspråket leta efter lämpliga motsvarande uttryck i källspråket. Att ta med typiska vändningar på målspråket och översätta till källspråket är i regel ingen bra lexikografisk metod, eftersom det lätt leder till otypiska ingångar, som ingen modersmålstalande användare kommer på tanken att slå upp. Men om man söker motsvarigheter till de typiska vändningarna i ett autentiskt material kan metoden ge en hel del matnyttigt.

Möjlighet till morfologiska sökningar i de svenska korpusarna

I svenska korpusar når man hyggliga resultat vid fritextsökning, eftersom orden med stamförändringar är få. Men ibland vore möjligheten till sökning på morfologisk grund önskvärd, särskilt när man arbetar kontrastivt med svenska och ett annat språk. Sökningar som finska språket aktualiserar och som förhoppningsvis så småningom blir möjliga att genomföra i Språkbankens korpusar är t.ex.:

- Att kunna hålla isär adverb och t-former av adjektiv och particip så att man vid en sökning får listat bara de belägg där -t-formen faktiskt är adverb.

- Att vid finalalfabetiska sökningar kunna söka på (efterleds)stam och inte som nu vara tvungen att göra en separat sökning för varje form av ordet, t.ex. *-flicka*, *-flickan*, *-flickas*, *-flickans*, *-flickor*, *-flickorna*, *-flickors*, *-flickornas*.

En annan sökning vi på den finsk-svenska ordboksredaktionen ofta saknat i Språkdatas korpusar är möjligheten att leta efter ordkombinationer där de två orden skiljs åt av ett eller flera andra ord. Med en sådan möjlighet kunde man

- söka efter prepositionsfraser och valensuppgifter i större omfattning än vad som nu är möjligt

- kontrollera semantiska selektionsregler – t.ex. hurdana typer av subjekt respektive objekt ett visst verb kan ta. Det är ju ofta en synnerligen viktig information i en tvåspråkig produktionsordbok.

Sådana sökmöjligheter finns inte åtminstone tills vidare i den version som finns tillgänglig över nätet och som vi på forskningscentralen utnyttjar.

En tvåspråkig korpus

Att skapa en omfattande korpus där samma texter ingår på båda språken är ett stort projekt. Dels är det naturligtvis dubbelt jobb att få fram samma texter på båda språken (dubbelt fler rättigheter att skaffa fram t.ex., eller rent av tre gånger fler om man också tar med översättningar från tredje språk). För det andra är det ju en dubbel uppsättning texter som skall kodas på lämpligt sätt.

Åtminstone i den första fasen kunde man därför skapa en mindre omfattande korpus som i första hand skulle utnyttjas för att undersöka

motsvarigheterna till vissa ordtyper, inte för enskilda ord. I den mån korpusen kunde utnyttjas för att kontrollera också översättningar av enstaka ord är allt gott och väl, men det skall inte vara det primära målet.

Tanken på en parallellkorpus för finska och svenska uppstod under arbetet med att redigera Finsk-svensk ordbok, eftersom vi under hela projektet gick en ojämn kamp med vissa oöversättbara ordtyper, ord som antingen helt och hållet saknar svenska ekvivalenter eller vars ekvivalenter inte kan ingå i samma typer av konstruktioner. Som lösning på problemet bestämde vi oss för att utarbeta syntaktiska modeller för hur orden kan hanteras vid översättning till svenska. (LexicoNordica 2:89 ff.). För att få fram exempelmaterial använde vi dels Helsingfors universitets finska korpus, dels ett par tre pocketböcker i en finsk och en svensk version parallellt. Den här parallelläsningen visade sig vara nyttig i all sin enkelhet och en mer omfattande korpus skulle säkert ge betydligt mer material.

Idén med modellartiklar har vi tänkt vidareutveckla i framtida upplagor. Dels kan de befintliga modellerna kanske ytterligare bearbetas utifrån korpusmaterialet, dels finns det fler ordtyper som vi gärna skulle åskådliggöra i modeller. Efter det att artikeln för LexicoNordica 2 skrevs har vi utarbetat en modell också för kausativa/faktiva verb. Den var vansklig att göra och skulle säkert vinna på en genomgång av korpusmaterial.

En typ vi gärna skulle komma åt är paren punktuella – iterativa/durativa verb som finskan är så rik på. Det punktuella verbet motsvaras ibland men långt ifrån alltid av partikelverb med *till* på svenska. Den iterativa typen är ännu mer problematisk. Finska iterativa verb används dels för att ange att det plötsliga skeendet sker gång på gång, dels tillsammans med plurala subjekt. Ibland kan svenskan upprätthålla en liknande skillnad med hjälp av ett adverbial av något slag i satsen, men ofta verkar det som den här aspekten inte skulle uttryckas på svenska. Hur den problematiken skall lösas i en ordbok är långt ifrån klart, men en parallellkorpus kunde ge ett bättre underlag för kartläggningen av problemet.

En annan typ av verb som kunde behandlas i en modellartikel är verbpar som består av ett intransitivt verb på *-ua/-yä* och ett motsvarande transitivt (t.ex. *muuttua* 'förändras' – *muuttaa* '(för)ändra'). Par av denna typ är mycket vanligare i finskan än vad par med -s-verb och transitivt verb är på svenska. (Se närmare Martola 1995a:287 f.)

En grupp av ord där den tvåspråkiga korpusen också kunde vara till hjälp är sammansättningar av bahuvrihityp. Det finns betydligt fler lexikaliserade bahuvrihisammansättningar på finska än på svenska. De

finska orden kan dessutom i regel vara både adjektiv och substantiv och som substantiv kan de ofta avse både egenskapen och bäraren av egenskapen.

Ytterligare en grupp som det vore intressant att undersöka med hjälp av en parallellkorpus är komparativ- och superlativformer av sådana ord vars svenska motsvarigheter inte kan kompareras. Den här typen är synnerligen besvärlig att behandla i en vanlig ordboksartikel (t.ex. *laidempi* 'som befinner sig närmare kanten' till substantivet *laita* 'kant, rand').

Problemet med en parallellkorpus är naturligtvis först och främst att uppbringa samma texter på bägge språken. Om man skall hitta samma ställe i båda texterna kan man inte använda sig av texter där den ena versionen är förkortad eller bearbetad, utan de borde i väldigt hög grad motsvara varandra.

Tanken är att i så hög grad som möjligt dra nytta av det finska korpusmaterialet, så att t.ex. de romaner som finns översatta tas med också på svenska. Många av de fenomen vi vill komma åt är relativt talspråksnära, så romaner utgör därför ett bra material, och också t.ex. barn- och ungdomsböcker. För att balans mellan språken skall nås bör det finnas med översättningar i bägge riktningarna och gärna också översättningar från tredje språk.

Sammanfattning

Svenska avdelningen vid Forskningscentralen för de inhemska språken har planer på dels en utvidgad finlandssvensk korpus, dels en tvåspråkig parallellkorpus. Den existerande finlandssvenska korpusen kommer att byggas ut i samarbete med Institutionen för nordiska språk vid Helsingfors universitet. Tills vidare har det konkreta arbetet inte inletts. I artikeln beskrivs främst de behov av korpusar som gjort sig gällande i arbetet på en finsk-svensk storordbok.

Litteratur

Tryckta källor

Finsk-svensk storordbok. Forskningscentralen för de inhemska språken. WSOY. Under tryckning.

Gellerstam, Martin 1986: Translationese in Swedish novels translated from English. I *Translation Studies i Scandinavia.* Ed. Lars Wollin and Hans Linquist. Lund: CWK Gleerup, 88–95

Lehtinen, Marja – Karvonen, Pirjo – Rahikainen, Tarmo 1995: *Tekstikorpukset*. Kotimaisten kielten tutkimuskeskuksen julkaisuja 81. Helsinki.

Martola Nina 1995a: Från enspråkig till tvåspråkig ordbok. I *Nordiske studier i leksikografi* 3. Skrifter utgitt av Nordisk forening for leksikografi, 283–293

Martola Nina 1995b: Substantiv avledda av adjektiv och verb – en jämförelse mellan några ordböcker med finska som källspråk. I *LexicoNordica* 2, 89–108

Otryckta källor

Sinclair, John 1992: Lexicographers' Needs. Pisa Workshop on Text Corpora. (stencil)

Le Parole. Technical Annex. Nov 5th 1995. Rev. Nov 19th 1995 - V1.5 (stencil)

Övrigt

Språkbanken, Göteborgs universitet

FISC-korpusen, Helsingfors universitet