Rapporter om svenska som andraspråk (ROSA) 8

OrdiL

En korpusbaserad kartläggning av ordförrådet i läromedel för grundskolans senare år

Redaktörer Inger Lindberg och Sofie Johansson Kokkinakis

Institutet för svenska som andraspråk
Göteborgs universitet

ROSA 8

OrdiL

– en korpusbaserad kartläggning av ordförrådet
i läromedel för grundskolans senare år

Redaktörer Inger Lindberg och Sofie Johansson Kokkinakis

ROSA utkommer oregelbundet. Serien består av forskningsrapporter m.m. inom ämnet svenska som andraspråk. Det främsta syftet med serien är att ge en möjlighet att snabbt och i preliminär form avrapportera arbetet inom Institutet för svenska som andraspråk. Även andra arbeten inom ämnesområdet publiceras dock. Frågor och synpunkter är välkomna och kan riktas direkt till författarna eller till Institutet för svenska som andraspråk, Institutionen för svenska språket, Göteborgs universitet, Box 200, 405 30 Göteborg.

TIDIGARE UTGIVNA RAPPORTER:

Anna-Britta Wallerstedt (1977) *Den receptiva ordförståelsen hos invandrarelever och deras inlärningssituation.*

Monica Reichenberg Carlström (1998) *Koheres, röst och läsning på ett andraspråk.*

Roger Källström (1999) *Svenska som andraspråk – lärarkompetens och lärarutbildningsbehov.*

Suzanne Nordin-Eriksson (under medverkan av Anna Kumlin) (2000) *Inlärarautonomi speciellt ifråga om lågutbildade andraspråksinlärare.*

Uno Källtén (2001) *Analys av Skolverkets rapporter och trycksaker under åren 1994 och 1999.*

Ulla Sundemo och Monica Nilsson (2004) *Barnboksfiguren – en tillgång på olika plan.*

Inger Lindberg och Karin Sandwall (red) (2006) *Språket och kunskapen – att lära på sitt andraspråk i skola och högskola.*

Institutet för svenska som andraspråk
Institutionen för svenska språket
Göteborgs universitet
Box 200
405 30 GÖTEBORG

Form: Conny Nylén, Mediavarvet

Reprocentralen
Humanisten
Göteborgs universitet

Innehållsförteckning

Förord

Ordförrådet i skolans läromedel kan vara en stor stötesten för många flerspråkiga elever. Att tillägna sig de ord och fraser som krävs för att förstå det som står i läroböckerna i olika ämnen kan lätt te sig som en oöverstiglig uppgift. Behovet av en effektiv och systematisk språkundervisning i detta avseende är därför stort. I denna rapport redovisas arbetet i ett projekt som syftar till att kartlägga ordförrådet i vanliga läroböcker i NO, SO och matematik för grundskolans senare år. Tanken är att en sådan inventering på olika sätt kan underlätta arbetet med att bygga upp ett adekvat, skolrelaterat ordförråd och vara till stor hjälp för både lärare och elever – inte minst när det gäller att veta vilka ord som är vanliga och viktiga i läroböckerna och därför nödvändiga att kunna.

Vi vill här särskilt tacka de förlag och författare som så generöst har ställt sina läroböcker i digital form till vårt förfogande. Utan detta samarbete hade projektet inte varit möjligt att genomföra. Vi vill också tacka Dimitrios Kokkinakis för betydelsefulla språkteknologiska insatser i projektet under flera år samt Karin Sandwall som under en period deltog i projektet med värdefulla didaktiska kunskaper och som dessutom har arbetat med utformningen av rapporten.

Göteborg den 5 februari 2007

Inger Lindberg
projektledare

OrdiL – en korpusbaserad kartläggning av ordförrådet i läromedel för grundskolans senare år

Syfte och bakgrund

Syftet med det projekt som presenteras i denna rapport är att utifrån en kartläggning av ordförrådet i vanligt förekommande läroböcker i grundskolans senare år framställa ordlistor med uppgifter om frekvens och spridning för ord i grundskolans ämnesundervisning. Utgångspunkten är att en fördjupad kunskap om lexikala drag som karakteriserar läromedel i olika skolämnen kan underlätta utvecklingen av ett skolrelaterat språk för elever med svenska som andraspråk.

En central del i projektet utgörs av uppbyggnaden av en läromedelskorpus, där läromedelstexter i NO- och SO-ämnen samt i matematik

omfattande cirka en miljon löpord samlas och görs tillgängliga och sökbara i digital form med hjälp av olika språkteknologiska verktyg. Genom läromedelskorpusen bidrar projektet med empiriska data även för en mängd andra framtida studier av autentiskt skolrelaterat språkbruk med varierande didaktiska tillämpningar.

En kartläggning av ordförrådet i ämnesundervisningens läroböcker kan bidra med viktig kunskap om lexikala aspekter av ett skolbaserat och skolrelevant språk. Det gäller inte minst för elever med svenska som andraspråk, för vilka ordförrådet i skolans ämnesundervisning utgör en särskild stötesten. Genom uppgifter om ordens frekvens och spridning kan ordlistorna ge såväl forskare, lärare som läromedelsförfattare viktig kunskap om de lexikala krav som läromedelstexter i olika ämnen ställer. Det är information som bl.a. kan läggas till grund för ett mer systematiskt, effektivt och ämnesövergripande pedagogiskt arbete med utvecklingen av ett skolbaserat ordförråd. En sådan kartläggning kan ge information om vilka ord utöver de mest frekventa som är vanligt förekommande i alla texter inom det skolrelaterade registret oberoende av ämne – ord som alltså skulle kunna betecknas som ämnesneutrala och allmänt skolrelevanta. Den kan även ge information om typiska fackord inom respektive ämnen liksom om fackspecifika användningar av mer frekventa och vardagliga ord.

Det är emellertid viktigt att påpeka att man här inte enbart ska fokusera på ordens frekvens och spridning. Genom djupare analyser av de ordens användning i olika kontexter t.ex. genom s.k. *konkordanser* (se vidare i kapitel 3 i denna rapport) kan man också få information om en rad olika aspekter på orden och deras användning i läroböckerna såväl i fråga om deras olika betydelser som i fråga om deras grammatiska mönster och kollokationer, d.v.s. i vilka frekventa ordkombinationer orden ingår.

Läromedelskorpusen och ordlistorna ger också ett utmärkt underlag för konstruktion av olika typer av diagnostiska bedömningsinstrument som kan användas för att mer systematiskt följa utvecklingen av elevernas receptiva och produktiva ordförråd ur såväl kvantitava

som kvalitativa aspekter. Flerspråkiga elever har ofta ett i många avseenden väl fungerande svenskt vardagsspråk som kanske inte på något uppenbart sätt avviker från enspråkiga elevers. Utan tillgång till evidensbaserade diagnostiska instrument för mätning av skolrelaterad ordkunskap kan det därför vara mycket svårt att fastställa om enskilda elever kan ha brister i fråga om centrala aspekter av ordförrådet och avgöra vilken typ av pedagogiska insatser som behöver vidtas. Det är här viktigt att påpeka att det faktum att en elev är flerspråkig naturligtvis inte alls behöver innebära att sådana brister föreligger. Avsaknaden av valida och reliabla instrument för diagnostisering av skolrelaterade språkfärdigheter är emellertid ett stort och allmänt omvittnat problem i skolan idag och bedömning av andraspråksfärdigheter sker ofta på godtyckliga grunder (Myndigheten för skolutveckling 2004). Det innebär att de pedagogiska insatserna på andraspråksområdet ofta baseras på ett bristfälligt underlag och kanske heller inte alltid når relevanta elevgrupper. Eftersom behovet av evidensbaserade diagnosinstrument med ett tydligt fokus på elevernas skolrelaterade språkfärdigheter är akut kommer utvecklingen av denna typ av instrument att prioriteras i det fortsatta arbetet med olika tillämpningar av projektet.

I denna rapport redovisas det arbete som genomförts i projektet fram till och med juni 2006. Här behandlas bl.a. det språkteknologiska arbetet med uppbyggnaden av läromedelskorpusen och utveckling av gränssnitt för sökning i korpusen, pågående arbete med olika typer av lexikologiska analyser, utprövning av modeller för bedömning av ordförråd samt i någon utsträckning även preliminära resultat i projektet. Rapporten innehåller en inledande forskningsbakgrund (kapitel 1) följd av en redogörelse för lexikologiska (kapitel 2) och språkteknologiska (kapitel 3) aspekter på projektarbetet. I de två följande kapitlen (4 & 5) redovisas olika didaktiska tillämpningar av projektets resultat med avseende på bedömning av ordförråd som genomförts i projektet. Slutrapportering av ordförrådsstudien beräknas ske under 2007.

1. Forskning om läromedelsspråk och ordförrådsutveckling

Inger Lindberg

I detta inledande kapitel ges en översikt över tidigare forskning kring läromedelsspråk, ordförståelse och ordförråd med särskilt fokus på aspekter som ansetts relevanta ur ett andraspråksperspektiv.

1.1 Lärandets språkliga dimensioner

1.1.1 Språk och skolframgång

Idag har nästan 15 % av alla elever i den svenska skolan en flerspråkig bakgrund. I storstäderna är denna siffra betydligt högre och i enskilda skolor kan i vissa klasser så gott som alla elever tillhöra denna kategori. Många av dessa elever har när de börjar i svensk skola (oavsett om det sker i år 1 eller senare) begränsade kunskaper i undervisningsspråket svenska och tvingas långt innan deras andraspråk är fullt utvecklat använda detta som instrument för lärande i skolans övriga ämnen och för sin allmänna kognitiva utveckling. Att detta i många fall

blir en alltför svår uppgift framgår av Skolverkets statistik som visar att elever med utländsk bakgrund i näst intill alla ämnen är starkt överrepresenterade bland de elever som inte når grundskolans mål. Angående resultaten av ämnesprovet i matematik från skolår 9 från 2004 skriver Skolverket t.ex. (SOU 2004:97, s. 237):

> Det förekommer stora skillnader mellan elever med svensk bakgrund och elever med utländsk bakgrund. 11,3 procent av eleverna med svensk bakgrund nådde inte betyget Godkänd i ämnesprovet i matematik, medan motsvarande andel för elever med utländsk bakgrund var 21,2 procent.

Även matematikdelegationen har påpekat att man särskilt behöver uppmärksamma elever med ett annat modersmål än svenska eftersom god språkförståelse är grundläggande för begreppsutvecklingen. Delegationen vill därför i sin handlingsplan ”ge lärare i förskola och skola möjlighet till kompetensutveckling i svenska och matematik ur ett andraspråksperspektiv.” (SOU 2004:97, s. 90, 137).

Lärandets språkliga dimensioner har bl.a. utforskats i australisk skolforskning i kretsen kring lingvisten Michael Halliday inom ramen för systemisk funktionell lingvistik (SFL) (Halliday 1994). Utgångspunkten för denna språkteori är språket som socialt fenomen och som verktyg för meningsskapande i olika sociala kontexter. Här lägger man också stor vikt vid språkets betydelse för lärande och skolframgång. I flera stora projekt har man t.ex. försökt tydliggöra kopplingen mellan de kunskapsmål som uppställs inom olika skolämnen och det språk och språkbruk man måste behärska för att uppnå dessa mål (se vidare Rothery 1996). Det har skett genom att man på olika sätt försökt kartlägga skolans läs- och skrivpraktiker, t.ex. genom att samla in elevtexter från en rad olika skolämnen och analysera dem med avseende på olika språkliga drag. Elevernas tillgång till relevanta modeller för de olika texttyper som anses centrala och representativa för olika ämnen har också undersökts genom att man inventerat läromedelstexter och annat undervisningsmaterial.

Lärande inom ramen för olika kunskapsområden i skolans ämnesundervisning förutsätter tillgång till och kontroll över ett specialiserat fackspråk genom vilket de idéer och begrepp som ligger till grund för ett visst kunskapsfält kan relateras till varandra och vävas samman till en meningsfull helhet. Det specialiserade språket är en semantisk resurs som möjliggör den typ av ”vetenskapligt” meningsskapande som kännetecknar mycket av undervisningen i skolan, inte minst i NO-ämnena. Det är också ett språk som fungerar som inträdesbiljett till den sociala gemenskap som utövare inom ett visst ämnesområde ingår i. Språket spelar alltså, precis som i den tidiga barndomen, en central roll även i den socialiseringsprocess som leder till medlemskap i de nya gemenskaper och roller som kemister, biologer, historiker o.s.v. som skolan ska erbjuda. Följaktligen ingår det i t.ex. historielärarens uppgifter att introducera och systematiskt träna sina elever inte bara i att tänka utan också att tala och skriva som en historiker (jfr Olvegård 2006).

1.1.2 Vardagsspråk och skolspråk

Skolarbetet innebär språkanvändning i alltmer kognitivt krävande aktiviteter och uppgifter, där språket ska fungera både som ett avancerat tankeredskap och som ett verktyg genom vilket kunskaper ska utvecklas, fördjupas och redovisas. Det ställer stora krav på elevernas förmåga att förstå och verbalisera abstrakta begrepp och komplexa samband. Dessutom ska kunskapen värderas ur ett kritiskt perspektiv vilket kan utgöra en särskild svårighet för andraspråkselever som ofta får lägga ner ett mycket större arbete än andra elever på den rent språkliga förståelsen av stoffet. Utvecklingen av de specialiserade språk som utmärker skolans ämnesundervisning tar dock tid och går inte av sig själv. Enligt nordamerikansk forskning (Thomas & Collier 1997) kan det ta mellan fem och tio år att utveckla den typ av skolrelaterade andraspråksfärdigheter som krävs för framgångsrika studier på grundskolans högstadium och i gymnasiet.

Språkutveckling under skoltiden handlar i stor utsträckning om erövringen av det skrivna språket. Det är ett språk som inte är modersmål för någon, men som skolbarn från olika sociokulturella miljöer i mycket varierande utsträckning kommit i kontakt med och förberetts för inför skolstarten. Som många lärare får erfara, kan man inte lita på att utvecklingen av det skriftspråksbaserade skolspråket sker automatiskt utan pedagogisk mediering. Det beror bl.a. på att barn och ungdomar från olika miljöer har högst varierande erfarenheter av de språkliga mönster som värderas och premieras i skolan. För elever från miljöer där hemmets språkanvändningsmönster överensstämmer med skolans kan det vara möjligt att utan särskild vägledning själva ”snappa upp” drag i det språk som kännetecknar t.ex. NO- och SO-ämnena och integrera det i sin egen språkliga repertoar. För andra elever, i synnerhet för dem som vuxit upp med begränsad kontakt med undervisningsspråket, innebär frånvaron av tydlig pedagogisk vägledning att de fråntas möjligheterna att vidga sin språkliga repertoar och utveckla språkliga verktyg för att förstå vad som kommuniceras i skolans ämnesundervisning och för att bli delaktiga i det meningsskapande som utmärker de natur- och samhällsvetenskapliga ämnena.

Gun Hägerfelth (2004) har i sin avhandlingsstudie om språkliga praktiker inom naturkunskapsämnet i gymnasiet tydligt visat att vissa elever betydligt lättare än andra lär sig skifta mellan vardagligt och naturvetenskapligt språk och också lägger stor vikt vid hur de ska formulera sig vetenskapligt i skrift. I följande exempel hjälps t.ex. Erik och Amer åt att formulera sig på ett ämnesrelevant sätt när de utför en uppgift kring energiutvinning genom kärnklyvning (Hägerfelth 2004:145):

Erik:	Den avger.
Amer:	En uranatom utger, sa du?
Erik:	Avger.
Amer:	Avger energi [skriver] genom genom att den
Erik:	den krockar med en neutron. Det är enklare att säga.
Amer:	Genom att den krockar med en neuntron. [skriver och läser upp]
Erik:	Du stryk *n* där. Genom att den krockar med en neutron delas …

Typiska skolrelevanta skriftspråksord (*avger*), passiva verbkonstruktioner (*delas*) liksom tydliga sambandsmarkörer (*genom att*) visar att Erik och Amer är på god väg in i det mer vetenskapliga skrivande som förväntas på i olika ämnen i gymnasiet. Andra elever, däremot, har svårt att överge ett vardagligt språkbruk. Samtalen stannar här ofta på en innehållsligt ytlig nivå som i följande exempel där några flickor diskuterar begreppet ***ekologisk nisch*** med forskaren/GH(Hägerfelth 2004:137):

Tina:	Jamen att till exempel att granar trivs bäst ute i skogen och att inte ute nånstans mitt mitt ingenstans, typ ute på gatorna eller nåt sånt där. De trivs ju bäst i skogen tillsammans med andra.
Astrid:	Ja med sina kompisar.
Tina:	Ja med sina kompisar.
GH:	Men varför gör dom det alltså. Varför?
Astrid:	Det är väl för att typ de lever i nåt särskilt klimat. Det kan vara luften där som de behöver. Äh, jag vet inte. Jag bara gissar.

Tina och Astrid visar här att de inte lyckas hantera begreppet *ekologisk nisch* inom ramen för den förväntade naturvetenskapliga diskursen. De baserar i stället sina förklaringar på en syn på t.ex. träd som bärare av mänskliga kvaliteter och känslor, ett synsätt som ibland också kan komma till uttryck i svensk naturskildringstradition men som knappast kan anses relevant i detta sammanhang. När man ska kommunicera naturvetenskap i skolan och uppmanas att förklara samspelet mellan olika organismer inom ett ekosystem förväntas man göra det inom ramen för en mer vetenskaplig diskurs där orden tilldelas en specialiserad naturvetenskaplig mening som omfattar en bestämd syn på hur naturen ska undersökas och förklaras.

För elever med andra modersmål än svenska, som inte har en lika trygg förankring i svenska språket som elever som har svenska som modersmål, kan övergången från vardagligt till ett ”vetenskapligt” språkbruk bli särskilt problematisk. Dessa elever, som ofta har sämre tillgång till modeller för skolrelaterat språkbruk utanför skolan, behöver därför en ämnesundervisning som är anpassad efter deras

särskilda behov och tar hänsyn till skolämnenas språkliga dimensioner utan att man ger avkall på kunskapskraven. För att dessa elever ska kunna nå de uppställda målen i skolans olika ämnen innan de fullt ut behärskar undervisningsspråket svenska krävs ofta också att de får ämnesrelaterad undervisning och studiehandledning på sitt modersmål. Ett viktigt krav på undervisningen för dessa elever är att den i största möjliga mån bidrar till att göra stoffet i de olika ämnena språkligt tillgängligt. Samtidigt bör den vara språkutvecklande och bidra till utvecklingen av elevernas allmänna såväl som skolrelaterade språkfärdigheter, inte minst när det gäller ordförrådet. Det förutsätter naturligtvis att alla lärare har kunskap om vari de språkliga kraven i samband med textläsning och andra språkliga aktiviteter i det egna ämnet består. I riktlinjerna för Lpo-94 (Skolverket 2006) anges också att varje lärare ska organisera och genomföra arbetet så att eleven får stöd i sin språk- och kommunikationsutveckling. Det innebär att elevens språkliga och kunskapsmässiga utveckling är en angelägenhet för alla lärare och skolledare.

1.2 Studier av läroboksspråk

1.2.1 Läsbarhetsformler

Lärobokstexter står av naturliga skäl i fokus för en hel del studier av språket i skolan. Inom denna forskning har man t.ex. studerat sambandet mellan olika språkliga drag och läsförståelse. Här dominerade länge, såväl i Sverige som internationellt, studier av ordförråd och läsbarhet. Olika formler för läsbarhet har utvecklats och tillämpats i studier där man utforskat sambandet mellan språkliga drag som t.ex. satslängd och individers resultat på lästest. Ett klassiskt verk i denna tradition utgör *A Teachers' Word Book of 10.000 words*, en engelsk frekvensordlista från 1921 (Thorndyke 1921) som utarbetats som vägledning för lärare. Lively och Pressey (1923) som har myntat begreppet *the vocabulary burden* hävdar med utgångspunkt från Thorndyke att läsbarheten i läroböcker sjunker ju fler lågfrekventa ord de innehåller. Så småningom har också olika formler som tar fasta på fler variabler som

anses påverka läsbarheten hos en text introducerats. En av de mest spridda är the *Reading easy formula* (Flesch 1949), en formel som bygger på satslängd och ordlängd. I Sverige introducerades på 1960-talet LIX (Björnson 1968) ett läsbarhetsindex som bygger på samma princip (se vidare i kapitel 4 i denna rapport).

Anne Charlotte Torvatn (2002) refererar i sin avhandling om textstrukturens inverkan på läsförståelsen i läroböcker en del av den kritik som riktats mot denna typ av läsbarhetsformler. Det har bl.a. påpekats att långa ord inte nödvändigtvis är okända eller svåra ord (se även kapitel 4 i denna rapport). Många långa ord är ju välkända (*chokladkaka*) medan många korta ord är relativt ovanliga (*stäv*). Vidare tas i dessa formler ingen hänsyn till om orden förekommer upprepade eller enstaka gånger i texten. Ett långt och ovanligt ord som förekommer upprepade gånger i texten blir rimligtvis lättare att tolka ju fler gånger man stöter på det. Dessutom har det påpekats att satslängd är ett problematiskt mått på grammatisk komplexitet eftersom det inte tar hänsyn till om satserna ingår i längre och mer komplexa satskomplex, vilket borde ha betydelse för läsbarheten. Själv kritiserar Torvatn det faktum att denna typ av formler värderar läsbarheten hos texter oberoende av ämne och innehåll och av vem som läser dem. Läsaren och dennes bakgrundskunskaper är alltså faktorer som man inte tar någon hänsyn till i detta sammanhang. Torvatn fann i sin avhandling att textstrukturen har stor inverkan på elevernas förståelse av lärobokstexter och att eleverna lättare kan ta till sig texter vars innehåll underbyggs med exempel och argument. Hon fann också att eleverna räknade med att definitionen av ett begrepp återfinns i texten efter omnämnandet av själva begreppet, vilket långt ifrån alltid var fallet.

I ovan nämnda läsbarhetsformler utgår man alltså från att kortare satslängd ger en mer läsbar text – ett antagande som starkt ifrågasätts i en svensk studie av läroboksspråk av Sylvia Danielson (1975) där kvantitativa aspekter på ordförråd och syntax i läromedel studerats. Danielson menar att korta satser i själva verket ofta utgör ett hinder för läsförståelsen (1975:131).

> När det gäller lärobokstext, kan man kanske peka på en annan orsak till ett samband mellan mycket korta meningar och svårlästhet. I texter med mycket korta meningar kan man finna textavsnitt, där de olika detaljerna i beskrivningen […] liksom staplas på varandra utan att man klart kan se sådana samband som förutsättning-följd eller orsak-verkan. Sådana samband är svåranalyserade, men icke desto mindre torde de spela en väsentlig roll, kanske framför allt i facktext. De alltför korta meningarna skulle med andra ord kunna ses om symptom på brister i den logiska uppbyggnaden av innehållet.

1.2.2 Nominal stil

En annan tidig svensk kvantitativ studie av lärobokstexter genomfördes av Margareta Westman (1974) inom projektet Bruksprosa. Här studerades texter från gymnasieböcker i biologi, samhällskunskap, historia, religionskunskap, psykologi och socialkunskap bl.a. utifrån aspekter som ordklassfördelning, ordförråd och meningsbyggnad. Sammanfattningsvis visar resultaten att läroboksspråket var koncentrerat och faktamättat och dessutom kännetecknades av en opersonlig och nominal stil där substantiv och adjektiv dominerar och verb och adverb används mer sparsamt. Ordförrådet kännetecknades av en relativt hög andel lågfrekventa, innehållstäta, inte sällan långa ord medan syntaxen var relativt enkel med förhållandevis få bisatser. Trots det är meningarna ofta långa till följd av att informationen pressats samman i mycket omfångsrika nominalfraser av typen ”ett exempel på en högst organiserad samhällsbildning med stark differentiering mellan individerna” som hämtats från en biologilärobok i undersökningen (Westman 1974:223).

1.2.3 Sammanhang i texten

På senare år har studier av läroboksspråk kännetecknats av en mer textlingvistisk inriktning där fokus har legat på faktorer som bidrar till att texten upplevs som en sammanhängande helhet snarare än en räcka lösryckta meningar. Ulla Ekvall (1995) har t.ex. i en studie

studerat begripligheten i lärobokstexter i orienteringsämnen för år 4 utifrån textlingvistiska analysmodeller, psykolingvistiska experiment, rekommendationer i skrivhandböcker och stilistiska kvaliteter. Ekvall fann många svagheter i texterna och påpekar bl.a. bristen på sammanhang och svårigheter med att ordna texternas delar i en hierarkisk innehållslig struktur. Anknytningen till det övergripande ämnet var ofta svag och oklar. Dessutom skymdes huvudtanken i texterna inte sällan av ovidkommande och perifera detaljer. Rubrikerna, som kan ha stor betydelse för att underlätta läsförståelsen genom att leda läsaren in i texten, lämnade också mycket övrigt att önska. Alltför snäva eller vida rubriker i förhållande till det påföljande innehållet innebär att de har ett begränsat värde för läsaren som stöd för den fortsatta läsningen.

I samma tradition har Lars Melin (1992) undersökt referentbindning, konnektion och tematik i lärobokstexter. Han påpekar bl.a. den ofta svaga kopplingen mellan olika meningar och begrepp som leder till att läsaren lämnas i sticket när det gäller att förstå hur olika enheter hänger ihop. Melin menar t.o.m. att texterna ofta är utformade som om innehållet vore känt för läsaren och att man för att förstå texterna alltså redan skulle behöva känna till innehållet, vilket knappast kan anses idealiskt i ett läromedel.

I en norsk studie påpekar även den norska andraspråksforskaren Anne Hvenekilde (Hvenekilde 1986) problemet med avsaknaden av explicita markörer för samband i texter, i synnerhet för elever med inlärningssvårigheter och för andraspråksläsare. I en analys av s.k. lättlästtexter fann Hvenekilde att de ändringar som genomförts för att göra texterna lättare ofta innebar att man gjort meningarna kortare. Detta ledde i allmänhet också till att de sambandsmarkörer som ger läsaren information om de logiska relationerna mellan meningarna togs bort, vilket avsevärt försvårar läsningen.

1.2.4 Kausalitet och röst

Att explicita markörer för orsakssamband kan påverka förståelsen av lärobokstexter i historia och samhällskunskap för elever med svenska som andraspråk framgår av en studie av Monica Reichenberg (2000) där hon bl.a. tydliggjorde orsakssamband genom att sätta in kausala satskonnektorer som *därför* och *därför att* i två autentiska texter. Reichenberg experimenterade också med att ge texterna *röst,* d.v.s. förse texterna med vissa talspråkliga drag för att minska distansen mellan författare och läsare. Denna bearbetning innebar t.ex. användning av aktiv form i stället för passiv, fullständiga satser i stället för satsförkortningar, korta fundament i stället för långa, mer frekventa i stället för mindre vanliga ord. Dessutom försågs texterna med en personlig och synlig berättarröst som vänder sig till läsaren med frågor och uppmaningar och tilltalar läsaren med ett direkt *du*.

Reichenberg skapade på så sätt tre nya bearbetade versioner av de två originaltexterna, en med mer kausalitet, en med mer röst och en med såväl mer kausalitet som röst. De olika textversionerna prövades sedan ut på 833 elever i skolår 7, varav ca hälften utgjordes av elever med svenska som andraspråk. Resultaten visar att bearbetning för röst respektive kausalitet och röst hade en särskilt gynnsam effekt på andraspråkseleverna som hade betydligt lättare att förstå dessa versioner jämfört med originalversionerna. Bearbetning med enbart kausalitet gav däremot inga signifikanta skillnader. Tilläggas bör att de bearbetade texterna i samliga fall var betydligt längre än originalversionerna och att kausalitet- och röstversionerna, d.v.s. de som eleverna hade lättast att förstå, också utgjorde de längsta texterna, vilket i sig kan ha påverkat förståelsen. Studiens resultat är därför svårtolkade.

1.2.5 Multimodalitet

Lars Melin (1995) har även i en studie uppmärksammat brister i läroböckers grafiska utformning. Böckerna ger enligt Melin ofta ett

splittrat intryck och kännetecknas av en svag och otydlig koppling mellan brödtext och bild/bildtext. Som flera andra forskare har påpekat (se t.ex. Hägerfelth 2004, Laursen 2004) är de visuella illustrationerna i dagens multimodala läromedel heller inte alltid valda för att förtydliga, konkretisera eller illustrera de verbala texterna. I stället har de ofta valts för att ge en estetisk upplevelse eller väcka elevernas intresse och motivation vilket innebär att sambandet mellan illustration och text ofta är långt ifrån självklart och entydigt.

1.2.6 Abstraktion

Agnes Edling (2004, 2006) har i en studie analyserat 58 läromedelstexter i svenska, SO och NO från skolår 5 och 8 i grundskolan och från gymnasiets andra år med avseende på abstraktionsgrad och funnit tydliga skillnader mellan ämnena. Hon utgår i sin analys från följande kriterier på konkreta och abstrakta ord/begrepp:

Konkret	**Abstrakt**
gripbart	inte gripbart
enskilt	inte enskilt
litet begreppsomfång	stort begreppsomfång
existerar i rummet och tiden	existerar inte i rummet och tiden
vardagligt	inte vardagligt

Edling nämner *flintastek* som exempel på ett begrepp som uppfyller alla ovanstående kriterier på ett konkret begrepp och *liberalism* som ett typiskt exempel på ett abstrakt begrepp.

Analysen visade att svenskämnets texter, som alla var berättande, innehöll högst andel konkreta och lägst andel abstrakta substantiv medan de naturorienterande texterna hade en betydligt lägre andel konkreta och en betydligt högre andel abstrakta och generaliserande substantiv. De samhällsorienterande texterna utgjorde en mellangrupp. Man kunde också se en stigande abstraktions- och generaliseringsgrad med stigande skolår i alla ämnen.

Men att förenkla texterna genom att göra dem mer konkreta är enligt Edling ingen lösning. Till skillnad från det vardagliga och mer verklighetsnära lärandet kännetecknas det specialiserade lärandet i skolan och i högre utbildning senare i livet av abstraktion och teknikalitet. Om man endast försöker omvandla abstrakta skoltexter till mer konkreta och vardagsnära texter i syfte att överbrygga klyftan mellan vardagskunskap och skolkunskap gör man därför enligt Edling eleverna en björntjänst. Den specialiserade skolkunskapen förutsätter nämligen ett annat språkbruk än den vardagliga kunskapen och låter sig inte uttryckas i ett vagt och oprecist vardagligt språk. Eleverna måste därför, menar Edling, ges möjlighet att successivt tillägna sig ett alltmer vetenskapligt språk, genom vilket de kan utveckla den typ av tekniska och abstrakta kunskaper som kännetecknar skolämnena på högre stadier.

Detta resonemang stämmer väl med Lev Vygotskijs (Vygotsky 1978, Vygotskij 1999) syn på språkets viktiga roll i utvecklingen av vetenskaplig förståelse, inte minst som verktyg för generaliseringar, kategoriseringar, abstraktioner, argument och reflektioner av den typ som utmärker lärandet i skolans senare stadier. Konstruktionen av det som Vygotskij kallat *vetenskapliga begrepp* kräver till skillnad från bildandet av spontana vardagliga begrepp en abstrakt förståelse som förutsätter utvecklingen av högre mentala funktioner. I denna process spelar språket en avgörande roll eftersom det är genom speciella sätt att använda språket de världsbilder konstrueras som ligger till grund för tankemönstren inom olika verksamheter. Man talar ofta om olika *diskurser* när man refererar till de språkbruk, betraktelsesätt och begrepp som styr tänkandet och vetenskapandet inom olika discipliner, praktiker och miljöer, som t.ex. inom naturvetenskapen. I stället för att undvika abstrakta texter i skolan bör man alltså ge eleverna rika tillfällen och möjligheter att röra sig i olika textvärldar. Det vardagliga och det vetenskapliga språkbruket står inte i motsättning till varandra. De utgör inte heller mer eller mindre betydelsefulla sätt att använda språket på. Det handlar i stället om *olika* sätt att använda språket på. För skolans elever innebär erövringen av det naturvetenskapliga språket kontakt med en helt ny språkkultur, vilket medför krav på olika typer av språklig anpassning. I mötet med skolans specialiserade

språk är det därför viktigt att eleverna inte uppfattar det som att deras vardagliga språkbruk inte duger eller måste korrigeras. Det handlar i stället om en utvidgning av deras språkliga repertoar.

1.2.7 Ordförråd i läromedel

Den ordförrådsstudie som presenteras i denna rapport har inspirerats av flera korpusbaserade ordförrådsstudier. De senare årens forskning inom korpuslingvistik har visat att texter i olika register, kännetecknas av olika språkliga drag, inte minst lexikala sådana (Biber 1989, Biber, Conrad & Reppen 1998). I en norsk undersökning (Golden & Hvenekilde 1983) kartlades ordförrådet i 40 läromedel i ämnena historia, fysik och geografi i år 4-9 i grundskolan. Syftet med studien var att utveckla övningsmaterial för andraspråkselever. Genom frekvensundersökningar av ordförrådet i dessa läromedel kom man fram till mycket intressanta och delvis överraskande resultat:

- Nästan hälften av orden i de enskilda läromedlen förekom endast en gång i läromedlet. 93 % av fackorden förekom bara i ett av ämnena. En del av dessa fackord är naturligtvis nya också för de enspråkiga eleverna och är alltså sådana ord som lärarna normalt förklarar för eleverna.
- 60 % av orden förekom i samtliga läromedel och utgjordes av 176 högfrekventa ord som man kan utgå ifrån att andraspråkseleverna kunde.

Den intressanta ordgruppen utgjordes dock av ord som varken tillhör de lågfrekventa eller de högfrekventa orden. Det är i stället ord som inte är så vanliga att man kan räkna med att andraspråkselever behärskar dem men som ändå är centrala för förståelsen av ämnesundervisningens texter. Dessa ord som Hvenekilde och Golden kallade *icke-fackliga ord* visade sig överraskande nog också till stor del vara ämnesspecifika.

- 55 % av dessa ord som t.ex. gnida i fysik och kvist i geografi förekom bara i ett av ämnena.
- Exempel på icke-fackliga ord som förekom i alla ämnen är behandla och krav.
- Geografi- och historieläromedlen visade sig innehålla betydligt fler s.k. icke-fackliga ord än t.ex. fysikläromedlen som för övrigt hade den största andelen fackord. 1/3 av orden i fysikläromedlen var fackord mot 1/6 i geografi och historia.

Golden och Hvenekilde menar att det är de s.k. *icke-fackliga orden* som vållar störst problem för andraspråkseleverna eftersom det finns stor risk att många lärare uppfattar dessa som kända för eleverna och därför inte förklarar dem närmare. Det kan alltså vara ord som vi inte direkt förknippar med ett speciellt ämne men som i skolans värld huvudsakligen förekommer i ett specifikt ämne. I annan relevant litteratur (Enström 2004) görs en distinktion mellan *ämnesspecifik* vokabulär och *allmänt akademisk* vokabulär där man med det senare avser ord som är vanliga i akademiska texter i samband med högre studier oavsett ämne. Det kan vara ord som t.ex. *ersätta*, *komplex* och *fastställa* och som ofta inte är särskilt framträdande i akademiska texter eftersom de sällan är centrala för de specifika ämnena i de texter där de förekommer.

1.3 Ordboksprojekt

1.3.1 The Academic Wordlist (AWL)

Ett tidigt och mycket ambitiöst korpusbaserat ordboksprojekt är den amerikanska ordboken *The American Heritage Schoolbook Dictionary* (Caroll, Davies & Richman 1971) som bygger på en korpus om drygt 5 miljoner löpord. Korpusen representerar alla typer av skolbokslitteratur som elever i den amerikanska skolan kan tänkas komma i kontakt med under år 3 – 9 och har 30.000 stickord och ca 70.000 de-

finitioner. En annan korpusbaserad ordlista som utkommit betydligt senare är *The Academic Word List* (AWL) (Coxhead 1998, 2000) som tagits fram av en forskargrupp vid Victoria University i Wellington, Nya Zeeland, i syfte att beskriva det akademiska ordförrådet i engelskan. Denna lista bygger på en korpus av engelskspråkig akademisk text i fyra discipliner och 15 ämnesområden på 3,5 miljoner löpord där frekvens och spridning hos ord utöver de 2.000 vanligaste orden undersökts. Listan innehåller 570 ordfamiljer som tillsammans täcker ca 10 % av det totala antalet ord i akademiska texter mot bara 1,4 % av orden i en skönlitterär korpus av motsvarande storlek vilket tyder på att listans ord innehåller företrädesvis akademiska ord. AWL kompletterar en tidigare framtagen ordlista, *The General Service List* (West 1953), som utvecklats från en korpus med 5 miljoner ord med inlärare av engelska som andra-/främmandespråk och deras behov som utgångspunkt. Denna lista innehåller de mest frekventa och ”användbara” 2.000 ordfamiljerna i engelskan. Den täcker upp till 90 % av orden i skönlitterära texter och upp till 76 % av orden i den akademiska textkorpus som ligger till grund för AWL.

Den korpus som ligger till grund för AWL bygger på akademiska texter om 3,5 miljoner löpord från ämnen inom 28 olika ämnesområden. Följande textavsnitt är hämtade från en av ekonomitexterna i the Academic Corpus med ord från Academic Word List understrukna:

> Dating the turning points and duration of business cycles has long been associated with the construction of aggregate reference cycle indexes, and their associated leading, coincident and lagging indicators. This was along lines originally developed by Burns and Mitchell (1946), and subsequently by colleagues at the National Bureau of Economic Research (NBER), e.g. Klein (1990). More recently, identifying the turning points and duration of business cycles has been an important aspect of two further areas of business cycle research: the evaluation of theoretical and associated empirical business cycle models, e.g. King and Plosser (1994), Simkins (1994); and the analysis of the time varying characteristics of business cycles, e.g. Diebold and Rudebusch (1992), Watson (1994).
> http://language.massey.ac.nz/staff/awl/corpus.shtml.

En central och återkommande fråga i detta sammanhang är vad som ska betraktas som ett ord, en fråga som för övrigt behandlas mer ingående i kapitel 2 i denna rapport. Ett sätt är att gruppera ord i ordfamiljer, vilket man gjort vid framställandet av AWL. Motivet bakom detta tillvägagångssätt är att man med kunskap om grundläggande böjnings- och ordbildningsregler i ett språk utan större ansträngning kan förstå ord som är regelbundet böjda eller avledda från medlemmar av samma ordfamilj. I AWL definieras en ordfamilj som en stam plus alla nära besläktade affixerade former, vilket inkluderar alla böjningar och de vanligaste produktiva och regelbundna prefixen och suffixen. Familjen inkluderar dock bara affix som kan läggas till stammar som kan stå som fria former. *Specificera* och *speciell* skulle alltså inte anses tillhöra samma familj eftersom formen *spec* inte förekommer som fri form. I tabell 1.1 ges exempel på ordfamiljer i AWL (Coxhead 2000:218):

Tabell 1.1 Exempel på ordfamiljer i AWL med utgångspunkt från respektive huvudord.

concept*	**legislate**	***indicate****
concept	legislate	indicate
conception	legislated	indicated
concepts	legislates	indicates
conceptual	legislating	indicating
conceptualisation	*legislation**	indication
conceptualise	legislative	indications
conceptualised	legislator	indicative
conceptualises	legislators	indicator
conceptualising	legislature	indicators
conceptually		

*De kursiverade orden utgör de mest frekventa i varje ordfamilj.

Orden i AWL listas också i underlistor där de 570 ordfamiljerna delats in i 10 listor med 60 ordfamiljer i varje (med undantag för lista 10 med 30 ordfamiljer) listade efter fallande frekvens. Som framgår av tabell 1.2 (Coxhead 2000:228) täcker orden i den första underlistan 3,6 % av ordförrådet, vilket innebär att de förekommer på ungefär var fjärde sida i engelska akademiska texter.

Tabell 1.2 Täckning för ord i sublista 1-10 i AWL efter fallande frekvens.

Sublista	Antal ord	Täckning i den akademiska korpusen i %	Kumulativ täckning i %	Upprepnings-frekvens per antal sidor
1	60	3.6	3.6	4.3
2	60	1.8	5.4	8.4
3	60	1.2	6.6	12.3
4	60	0.9	7.5	15.9
5	60	0.8	8.3	19.4
6	60	0.6	8.9	24.0
7	60	0.5	9.4	30.8
8	60	0.3	9.7	49.4
9	60	0.2	9.9	67.3
10	30	0.1	10.0	82.5

I tabell 1.3 återfinns de tio första orden i en alfabetiskt ordnad lista över huvudorden i AWL där numret på den frekvensordnade underlista som ordet och dess familjemedlemmar hör hemma i återges i kolumnen till höger (jfr även kapitel 3 i denna rapport).

Tabell 1.3 De tio första huvudorden i AWL med hänvisning till respektive frekvensordnad sublista.

Huvudord	Sublista
abandon	8
abstract	6
academy	5
access	4
accommodate	9
accompany	8
accumulate	8
accurate	6
achieve	2
acknowledge	6

1.3.2 LEXIN-projektet

I Sverige startades 1978 det s.k. LEXIN-projektet, ett forsknings- och utvecklingsprojekt som hade till syfte att utreda och skapa förutsättningar för produktion av lexikon för invandrare. LEXIN-projektet genomfördes i samarbete med dåvarande Statens institut för läromedel (SIL) och Skolöverstyrelsen (SÖ) samt flera universitets-institutioner i Sverige, däribland Göteborgs universitet som genom Språkdata tog fram den svenska ordbasen (se vidare Gellerstam 1978). LEXIN-projektet gick i första hand ut på att ta fram bra lexikon för invandrare i olika språkgrupper. Den svenska ordbok som fungerar som utgångspunkt för detta arbete (Skolöverstyrelsen 1985) bygger huvudsakligen på källor av (1) allmänt inriktade, icke textanpassade ordlistor, (2) ordförteckningar till läromedel för invandrare, (3) ordförråd inom området samhällsinformation bl.a. hämtat från ordlistor för tolkar. Ordurvalet för dessa lexikon bygger alltså inte på någon speciell korpus utan har en godtycklig sammansättning från källor som ansetts relevanta i sammanhanget. Fördelarna med korpusbaserade ordlistor är att de bygger på autentiskt material definierade till register, ämne och stil. Eftersom textunderlaget för sådana ordlistor numera ofta lätt kan göras tillgängligt i digitalt format kan en sådan insamling numera också utan större problem ske i stor skala. Den svenska LEXIN-ordboken finns idag med inspelat uttal av uppslagsorden tillsammans med versioner till elva olika språk på nätet på http://www-lexikon.nada.kth.se/skolverket/lexin.shtml

1.3.3 Nusvensk frekvensordbok (NFO)

Den utan jämförelse mest omfattande frekvensundersökningen av nusvenskans vokabulärsystem ligger till grund för Nusvensk frekvensordbok 1 – 4 (Allén 1970; Allén, Berg, Järborg, Löfström, Ralph & Sjögren 1980) och Tiotusen i topp (Allén 1972) som bygger på ett pressmaterial från 1965. Dessa ordböcker utgör mycket värdefulla jämförelsematerial för den här aktuella studien ur många olika aspekter. De kan dock knappast användas för att i likhet med *The General Service List* (West 1953) ge information om de mest frekventa

och ”användbara” orden för denna typ av texter eftersom de bygger på en korpus av tidningstexter. Liksom The General Service List är urvalets relevans också tveksamt med tanke på materialets ålder.

1.4 Studier av ordförståelse och ordförråd

1.4.1 Ordförståelse i förstaspråket

En tidig svensk ordförrådsstudie med relevans i detta sammanhang genomfördes under 1970-talet då mer än 700 ord med betydelse för det sociala, politiska och ekonomiska livet undersöktes med hänsyn till begriplighet i en studie som redovisades i boken Språkklyftan (Frick & Malmström 1976). Här tillfrågades cirka 700 deltagare i Arbetsmarknadsstyrelsens kurser vid sju AMU-centra om hur de uppfattade betydelsen hos ord på en lista med 100 ord som ansågs aktuella för en ”svensk i sociala, vardagsekonomiska, politiska och fackliga sammanhang” som man ansågs kunna möta i ”radio/TV, tidningarnas ledare, och på nyhets- och debattsidor” (ibid.:10). Exempel på sådana ord är *akutmottagning, anföra besvär, bordlägga, civilstånd, detaljhandel, fusion, produktivitet, soliditet* och *utanordna*. Många av orden kan hänföras till ett mer allmänt språkbruk som adjektiv och adverb av typen *aktiv, ambitiös, förkastlig, godtycklig* och substantiv som *ambition, aspekt, citat, dimensioner, tendens* och *teori* liksom verb som *acceptera, generalisera, hävda* och *improvisera*. Avsikten var att mäta det passiva ordförrådet, d.v.s. hur man uppfattar orden när man läser eller hör dem och inte hur de tillfrågade själva använde dem.

Urvalet gjordes utifrån ordboken *Ord vi möter* där en tredjedel av orden återfinns i åtta rapporter till Metallarbetarförbundets kongress 1973 medan andra ord är hämtade från brev, upplysningsskrifter eller blanketter som myndigheter, banker, och försäkringsbolag använder. Ytterligare ord hämtades från t.ex. tidningsledare, tidningsartiklar, politiska broschyrer, årsberättelser och personaltidningar. Specialtermer lämnades utanför undersökningen men ambitionen var att få med ord som måste förklaras ”för en inte alltför obetydande del av vad vi kan kalla ”vanliga” människor” (ibid.:11). I undersökningen

ansågs ord strax ovanför gränsen till de ”lätta” orden vara viktiga att pröva liksom ord som kunde antas ligga strax under denna gräns. Därutöver testades ett antal ord som antogs vara mycket svåra. Exempel på sådana ord var *acklamation, assimilation* och *juridisk person*.

Orden presenterades i meningar som vara avsedda att ge en föreställning om i vilka sammanhang och tillsammans med vilka andra ord som testorden brukar förekomma. Orden *drastisk*, *bordläggas* och *arvode* testades t.ex. i följande meningar.

Man tog till *drastiska* åtgärder.
Frågan bör *bordläggas*.
Arvodet är litet.

De tillfrågade fyllde i ett frågeformulär där de uppmanades kryssa för ett av alternativen ”Vet säkert vad ordet betyder”, ”Tror att jag vet vad ordet betyder” eller ”Vet inte vad ordet betyder” och därefter ombads att ange vad ordet betyder. Denna metod som kräver att de tillfrågade själva finner ett uttryck med samma betydelse som testordet ansågs ställa vissa krav också på den aktiva språkbehärskningen även om testet i huvudsak inriktades på den passiva ordförståelsen.

Resultaten visar att de undersökta orden förstås på ungefär följande sätt:

70 % anser sig förstå ordet.
30 % förklarar att de inte förstår ordet.
55 % av alla svar ger översättningar som bedöms som godtagbara.
5 % ger översättningar som bedömts som delvis rätt.
10 % kommer med klara feltolkningar.

Bland feltolkningarna urskiljs följande grupper:

1. feltolkningar styrda av ljudlikhet
   (egalt→legalt, förorda→förordna, delegera → dela upp)
2. förväxling med ord som har motsatt betydelse
   (debiteras→få betalt, intäkter→ betalningar, hypotes→slutsats)
3. annan förväxling med ord som berör samma område
   (komfort→lyx, amortering→skuld, häleri→stöld)
4. svårigheter att hålla isär ord eller olika betydelser av ett och samma ord
   (deklarera åsikter→deklarera inkomster, realisera planer→realisera varor,
   utgå (om förmån) →utgå (ur t.ex. ett lag)
5. svårigheter med ordsammansättningar
   (affärsbank→bank i affärer).

Tabell 1.4 Exempel på ord som vållade särskilt många av de tillfrågade problem (Frick & Malmström 1976).

Testord	**Andel ”vet inte”, felsvar eller ”delvis rätt” i %**
borgenär	91
verifikationer	75
delegera	85
mandat	91
reaktionär	85

De slutsatser som dras från denna undersökning är att en mycket stor del av de tillfrågade som man ”med viss rätt” ansåg representera svenska folket inte tillfredsställande kunnat förklara åtskilliga ord som är viktiga för vardagsekonomi och vardagsjuridik, läkarvård, fackligt arbete, föreningsliv och annat samhällsliv. Det kan medföra att stora grupper hindras från att utnyttja sina rättigheter och fullgöra sina skyldigheter i samhället. Det bör slutligen tilläggas att personer med högre utbildning var underrepresenterade bland de tillfrågade och att endast personer med svenska som modersmål deltog i undersökningen.

1.4.2 Ordförrådet i andraspråket

Ordförråd och läsning

Det är väl känt att storleken på läsarens ordförråd har stor inverkan på hur väl man kan tillgodogöra sig och ta till sig innehållet i en text (Read 2000). Ordförrådet anses också vara den enskilt viktigaste faktorn för skolframgång för den som studerar på sitt andraspråk (Saville-Troike 1984, Laufer 1996). Förståelsen av skolans läromedelstexter påverkas i hög grad av kunskap om de enskilda orden i texten och internationell forskning har visat att åtminstone 95 % av orden bör vara kända för att man med någorlunda behållning ska kunna läsa och ta till sig innehållet i en text (Nation 2001). Vissa forskare (Carver 1994) menar t.o.m. att mötet med mer än 2 % okända ord i en text blockerar förståelsen för andraspråksläsaren. En högre andel okända ord i en text påverkar läsförståelsen markant, inte minst genom att möjligheterna att sluta sig till okända ords betydelse utifrån kontexten avtar, något som generellt är svårare för en andraspråksläsare. Men läsarens tolerans för okända ord varierar naturligtvis också med andra faktorer, som t.ex. med texttypen och läsarens förtrogenhet med innehållet i texten.

Enligt en rad olika skandinaviska läsundersökningar (Fredriksson & Taube 2001, Skolverket 2003, Kulbrandstad 2003) uppvisar andraspråkselever också som grupp sämre läsprestationer än förstaspråkselever mycket beroende på ett mindre utvecklat ordförråd. Det beror delvis på skillnader i fråga om det receptiva ordförrådet som ofta är mer begränsat hos flerspråkiga elever eftersom de i lägre utsträckning kommit i kontakt med undervisningsspråket i mer skriftspråkliga register. Läsning anses ha särskilt stor betydelse för ordförrådets utveckling eftersom det skrivna språket har en högre lexikal täthet[1] än talat språk i vardagliga situationer (Huckin, Haynes & Coady 1993).

[1] Ett mått som används för att ange den procentuella andelen lexikala ord (ej funktionsord) i en text. Måttet beräknas: LD= ((Antal lexikala ordförekomster x 100)/totala antalet ord).

Det innebär att det skrivna språket har en förhållandevis hög andel innehållsord (substantiv, adjektiv, verb etc.) medan det talade språket kännetecknas av relativt fler funktionsord (pronomen, prepositioner etc.) Även om inlärning av nya ord i stor utsträckning sker s.a.s. oavsiktligt i samband med läsning är sannolikheten att man ska lära sig ett nytt ord i samband med läsning relativt låg (Swarnborn & de Glopper 1999). Det innebär att man bör möta ett ord ganska många gånger i en text för att chansen att man ska lära sig det ska vara någorlunda stor. Sannolikheten för att man ska lära sig ett ord vid första mötet är dessutom generellt lägre för unga läsare, i svåra texter och för läsare utan särskild träning i att dra slutsatser om okända ords betydelser utifrån kontexten (Carlo et al. 2004).

David Qian och Mary Schedl (2004) som har undersökt förhållandet mellan ordkunskap och läsförståelse och bedömning av läsfärdighet i engelska som andraspråk har funnit att såväl ordförrådets bredd som djup är av stor betydelse i detta sammanhang (se vidare i avsnitt 1.7.2 nedan). Två aspekter av ordförrådets djup, nämligen kunskap om ord med närliggande betydelser, *synonymer,* och kunskap om frekventa ordkombinationer i vilka ordet ingår, *kollokationer,* spelar dock enligt Qian & Schedl en särskilt viktig roll. Inte mindre än 60 % av variationen i lästestens resultat i en av hans studier kunde förklaras med skillnader relaterade till ovanstående variabler. Flera studier visar också på klara samband mellan bredd och djup i ordförrådet och att mått på ordförrådets storlek utifrån frekvens även kan ge en bra indikation på mer generell språkfärdighet (jfr t.ex. Zareva, Schwanenflugel & Nikolova 2005).

Kvantitativa och kvalitativa aspekter

Skillnader mellan ordförrådet i förstaspråk och andraspråk kan beskrivas både kvantitativt – utifrån ordförrådets storlek eller bredd – och kvalitativt – utifrån ordförrådets djup. Ordförrådets storlek hos enspråkiga barn vid skolstarten anses ligga på 8.000 – 10.000 ord och expansionen under skolåldern är kraftig med ca 3.000 ord/år under gynnsamma omständigheter (jfr Viberg 1993). Undersökningar

visar att tvåspråkiga barns ordförråd i majoritetsspråket såväl vid skolstarten som långt upp i skolåren ofta är avsevärt mindre omfattande än hos jämnåriga enspråkiga elever (se t.ex. Verhoeven & Vermeer 1985). I Danmark har Jürgen Gimbel (1997) genomfört en mindre undersökning av ordförrådet hos tvåspråkiga elever med turkiska som modersmål och enspråkiga danska elever i klass 5. De 50 ord som undersöktes var sådana ord som inte var rena fackord men som kan förknippas med olika skolämnen och med stor sannolikhet förekommer i lärobokstexter i dessa ämnen även på lägre stadier. Exempel på ord som testades är *ansvar*, *appetit*, *bønder*, *døgn*, *energi,* och *flod*. Gimbels undersökning visar att det kan finnas stora kvantitativa skillnader mellan enspråkiga och tvåspråkiga elevers ordförråd i danska. Resultaten visar att de tvåspråkiga eleverna i genomsnitt behärskade 15 ord (minimum 3 och maximum 37) medan de danska eleverna i genomsnitt kunde 42 ord (minimum 35 och maximum 47). Gimbel kunde också påvisa intressanta kvalitativa skillnader med avseende på ordkunskapen mellan de olika grupperna. Enspråkiga elever uppgav nämligen ofta betydelsemässigt relativt närliggande ord då de skulle gissa betydelsen av okända ord medan de flerspråkiga eleverna i större utsträckning tycktes lita till fonetiska snarare än semantiska ledtrådar och uppgav ord som ljudmässigt liknande det efterfrågade ordet.

I många internationella studier har man också funnit andra typer av kvalitativa skillnader som tyder på att tvåspråkiga barn och ungdomar kan ha en förhållandevis grund kunskap om orden i andraspråket (Verhallen & Schoonen 1993, Vermeer 2001, Golden 2005). Det kan bl.a. innebära att de har färre och/eller andra typer av associationer förknippade med orden i sitt ordförråd än enspråkiga elever. Shidrokh Namei (2002), som i sin avhandlingsstudie använt associationstest (se även 1.7.2) för att studera ordförrådets organisation och utveckling hos tvåspråkiga elever, har bl.a. visat att ytlig kunskap om orden ofta ger associationer till andra ljudlika ord medan en djupare kunskap om ordens betydelse leder till associationer på semantisk grund, vilket vittnar om en högre grad av integrering av ordet i det mentala lexikonet, vilket bekräftar Gimbels resultat ovan.

Forskning visar också att även barn som mycket tidigt kommit i kontakt med andraspråket genom att de är födda i landet eller har anlänt i mycket unga år, långt upp i skolåldern kan ha svårare att tolka abstrakta ord och metaforiska betydelser av ord och uttryck än elever som har undervisningsspråket som sitt förstaspråk (Hene 2004, Golden 2005). I en norsk studie har Anne Golden (2005) undersökt förståelsen av metaforiska uttryck av typen *vara på jakt efter*, *bombardera med reklam* och *ett tungt argument* i norska samhällskunskapsböcker och funnit att denna typ av uttryck är betydligt svårare att förstå för elever med norska som andraspråk än för förstaspråkselever. Detta bekräftar resultaten i en läsförståelsestudie av den norska läsforskaren Lise Iversen Kulbrandstad (1996) där hon bl.a. fann att rubriken *Medaljens baksida* i en lärobokstext om Japan var problematisk för andraspråkseleverna eftersom de inte tolkade uttrycket i dess överförda betydelse. Det innebar att de inte kunde utnyttja rubriken för att bygga upp förväntningar om innehållet i texten och därmed gick miste om viktigt stöd för den övergripande förståelsen i sin läsning.

Ordförrådets utveckling

I svensk forskning har ordförrådets utveckling i andraspråket bl.a. studerats av Åke Viberg (2004) som i sin forskning kring verblexikonets utveckling visat på en kraftig överextension av *nukleära verb*, d.v.s. verb med mycket grundläggande betydelser som *göra*, *titta* och *gå* hos andraspråksanvändare i tidiga inlärningsstadier. Så småningom sker en successiv semantisk differentiering av verblexikonet i andraspråket som t.ex. kan innebära att verbet *gå* kompletteras med andra rörelseverb som *åka*, *fara* och *springa*. Viberg har emellertid också funnit att vanliga flertydiga verb med språkspecifika mönster av besläktade betydelser som t.ex. verbet *få* kan vara starkt underanvända i andraspråksinlärares språk. Detta är dessutom verb som ofta har en rad grammatiska betydelser som kan vara svåra att komma underfund med (*Får vi komma in? Nu får du sluta! Vi fick just veta det. Han fick oss att skratta*). Att tillägna sig hela den uppsättning av betydelser och användningsområden som infödda språkbrukare behärskar för denna typ av verb är enligt Viberg en mycket utdragen process som kan ta mer än fyra år, vilket var den tidsperiod under

vilken de flerspråkiga barnen i studien studerades. Nederländska studier visar också att högpresterande andraspråkselevers ordförråd kan vara jämförbart med flera år yngre lågpresterande enspråkiga elevers ordförråd ur såväl kvantitativa som mer kvalitativa aspekter (Verhallen & Schoonen 1993, Vermeer 2001).

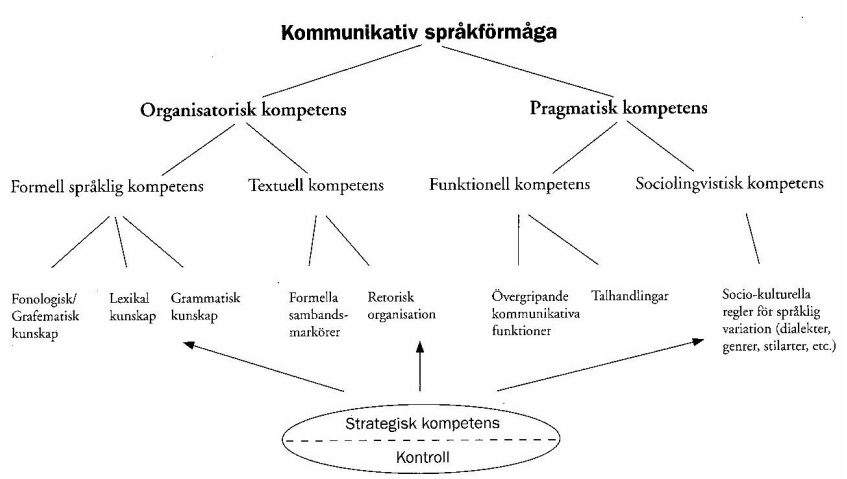
Kvalitativa brister i ordförrådet kan också få stor betydelse för andraspråkselevernas förmåga att tillägna sig kunskaper i skolans fackämnen men också för deras möjligheter att visa sina kunskaper på ett rättvisande sätt. Det faktum att elever använder ett ord i den egna produktionen behöver heller inte betyda att de behärskar ordets betydelse med alla dess konceptuella implikationer. Som Ing-Marie Parszyk (1999) visat kan det i matematikundervisningen t.ex. handla om rent matematiska begrepp som *area*, *omkrets, decimal* och *tredjedel* men lika gärna om helt vardagliga ord som i matematikundervisningen får en speciell betydelse. Verbet *uppskatta* har t.ex. en helt annan betydelse i en mer vardaglig kontext jämfört med när eleverna i en uppgift ombeds uppskatta en yta. Det kan göra en provuppgift obegriplig för många elever som kanske inte alls har några problem med själva matematiken. Som många andra forskare påpekat (jfr Hvenekilde 1991, Rönnberg & Rönnberg 2001, Norén 2006) är matematik ett ämne som även erbjuder andraspråkselever mer strukturella problem till följd av olika kulturers skillnader i talsystem, räkneord, och räknetraditioner.

1.5 Kommunikativ och lexikal kompetens i andraspråket

1.5.1 Kommunikativ kompetens

Frågan om hur den lexikala kompetensen förhåller sig till individens allmänna språkliga kompetens har intresserat många forskare (Reed 2000). Bachman och Palmer (1996, se även Lindberg 2004) beskriver en modell för kommunikativ språkförmåga indelad i *organisatorisk* kompetens och *pragmatisk* kompetens. Kunskap som relateras till det språkliga systemet på ord-, sats- och textnivå kallas här organisatorisk kompetens, medan förmågan att använda språket på ett funktionellt och sociokulturellt adekvat sätt i förhållande till olika syften, situationer, och genrer

hänförs till den pragmatiska kompetensen. Eftersom det knappast går att dra några skarpa gränser mellan grammatik och lexikon kan den lexikala kompetensen också ses som del av en mer övergripande och integrerad *lexikogrammatisk* kompetens där lexikon och grammatik samspelar på ett intimt och oupplösligt sätt genom att orden har en inherent egen grammatik med avseende på såväl morfologiska som syntaktiska drag.



Figur 1.1 Kommunikativ kompetens fritt efter Bachman & Palmer (1996).

1.5.2 Modeller för lexikal kompetens

Även om man kan anta att en högre allmän språkförmåga går hand i hand med en mer omfattande lexikal kompetens är det fortfarande oklart på vilket sätt inföddas lexikala kunskap skiljer sig från icke-infödda språkbrukares på olika färdighetsnivåer. Det faktum att den lexikala kompetensen är mångdimensionell och mångfacetterad innebär att det kan handla om skillnader av många olika slag. Att ”kunna” ett ord är inte en fråga om allt eller inget utan snarare ett graduellt fenomen som kan beskrivas som ett kontinuum från enbart igenkänning av ett ord till kunskap om alla aspekter av ett ords betydelse och användning i egen produktion. Sedan Jack Richards (1976)

presenterade en tidig modell för lexikal kompetens har flera forskare under senare år presenterat modeller för vilken kunskap om ett ord som krävs för en fullständig receptiv och produktiv behärskning. I svensk andraspråksforskning har Christopher Stroud (1979) presenterat följande modell över alla de aspekter som kunskap om ett ord inbegriper:

1) Ordens fonologiska uppbyggnad, d.v.s. deras ljudbilder.
2) Ordens morfologiska uppbyggnad, d.v.s. ordbildnings- och ordböjningsegenskaper.
3) Ordens kategoritillhörighet, d.v.s. vilken ordklass ett givet ord hör till.
4) Ordens syntaktiska egenskaper, d.v.s. vilka krav på frasens eller satsens övriga ord ett givet ord ställer.
5) Ordens betydelsemässiga egenskaper. Denna punkt innefattar:
    (a) Ordens semantiska uppbyggnad, vilka betydelsekomponenter som sammanfattas i ett givet ord.
    (b) Betydelserelationer mellan ord, t ex vilka motsatstyper som förekommer och vad det innebär att ett ords betydelse är överordnat ett annat ords betydelse.
    (c) Betydelserelationer inom ett ord, d.v.s. hur de betydelser som ett och samma ord kan ha är relaterade till varandra.
    (d) Ordens bibetydelser.
    (e) Ordens selektionsrestriktioner, d.v.s. vilka betydelser de ord får ha som ett givet ord kan kombineras med.

En annan modell som man ofta refererar till har presenterats av Nation (2001). Här ges en något annorlunda bild av den lexikala kompetensen som även inbegriper kunskap om ordens frekvens och kollokationella egenskaper.

1) Ordets talade form.
2) Ordets skrivna form.
3) Ordets grammatiska egenskaper.
4) Ordets kollokationella egenskaper.
5) Ordets frekvens.
6) Ordets stilistiska restriktioner.
7) Ordets konceptuella betydelse.
8) Ordets associationer till andra ord.

Senare modeller bygger på färre och mer globala aspekter (jfr även Henriksen 1999) som den modell bestående av fyra interagerande och sammanvävda dimensioner som introducerats av Qian (2002:516) och som också rymmer mer dynamiska perspektiv (min övers):

1) Ordförrådets storlek eller bredd, d.v.s. det antal ord som inläraren har åtminstone någon ytlig kunskap om.
2) Ordförrådets djup, som inkluderar alla lexikala egenskaper som fonologiska, grafematiska, morfologiska, syntaktiska, semantiska kollokationella och fraseologiska egenskaper liksom frekvens och register.
3) Lexikal organisation som refererar till lagring, kopplingar mellan och representation av orden i inlärarens mentala lexikon.
4) Automatisering av receptiv-produktiv kunskap med hänsyn till alla grundläggande processer genom vilka åtkomst till ordkunskape möjliggörs för såväl receptiva som produktiva syften, inklusive fonologisk och ortografisk kodning och avkodning, åtkomst till strukturella och semantiska drag från det mentala lexikonet, lexikosemantisk integrering och representation samt morfologisk analys och komposition.

Qian menar att dessa dimensioner är intimt och intrikat sammankopplade och interagerar i alla grundläggande processer som har med ordförrådets användning och tillväxt att göra. Betydelsen av enskilda drag varierar dock i t.ex. receptiva och produktiva processer.

1.5.3 Receptivt och produktivt ordförråd

Termerna *receptivt* och *produktivt ordförråd* används numera ofta synonymt med *passivt* och *aktivt ordförråd*. Som Nation påpekat (2001) är varken den ena eller den andra terminologin emellertid helt adekvat eftersom vi ”aktivt producerar” och konstruerar betydelse även när vi lyssnar och läser. Även om reception och produktion kan ses som ett kontinuum anser vissa forskare att denna distinktion också har att göra med de olika typer av associationer som är knutna till den aktiva respektive passiva vokabulären. Nation refererar här till Meara (1990) som menar att det aktiva ordförrådet kan aktiveras av andra ord medan ord som tillhör det passiva ordförrådet bara aktiveras genom att man ser eller hör deras form.

Corson (1995) som baserar distinktionen mellan aktivt och passivt ordförråd på användning snarare än på grad av kunskap menar att det passiva ordförrådet utöver det aktiva inkluderar ord som delvis är kända, lågfrekventa ord liksom ord som undviks i egen aktiv produktion. Vissa ord kan ju vara väl kända utan att därför komma till aktiv användning. Corson (1985) har introducerat begreppet *den lexikala barriären* som han menar vara avgörande för distinktionen receptivt/produktivt ordförråd. Han hänvisar i detta sammanhang särskilt till det mestadels lågfrekventa grekisk-latinska ordförrådet i engelskan som för många engelsktalande utgör ett passivt ordförråd. Corson menar också att detta ordförråd för de flesta talare av engelska som förstaspråk är mindre tillgängligt då de grekisk-latinska orden har en morfologisk struktur som avviker från den som kännetecknar det mer högfrekventa och vardagliga germanska ordförrådet. Corson menar att den barriär som skiljer dessa ordförråd stänger ute många engelsktalande från högre studier och också utgör en barriär mellan två vitt skilda betydelsesystem, d.v.s. mellan å ena sidan ett vardagligt och å andra sidan ett mer statusfyllt akademiskt betydelsesystem. Tillgången till den akademiska världens ordförråd och meningsskapande är därför enligt Corson en förutsättning för utbildningsframgång.

> What the lexical bar represents is a gulf between the everyday meaning systems and the high status meaning systems created by the introduction of an academic culture of literacy. This is a barrier that everyone has to cross at some stage in their lives, if they are to become 'successful candidates' in conventional forms of education. (ibid: 180-181)

Nation (2001) påpekar att distinktionen receptiv/produktiv kan tillämpas på olika typer och aspekter av ordkunskap och presenterar en modell som visar hur den lexikala kunskapen inkluderar både form, betydelse och användning som alla i sin tur kan relateras till såväl receptiv som produktiv kunskap.

Tabell 1.5 Vad innebär det att kunna ett ord? (Nation 2001:27)

Form	spoken	R	What does the word sound like?
		P	How is the word pronounced?
	written	R	What does the word look like?
		P	How is the word written and spelled?
	word parts	R	What parts are recognisable in this word?
		P	What word parts are needed to express the meaning?
Meaning	form and meaning	R	What meaning does this word form signal?
		P	What word form can be used to express this meaning?
	concept and referents	R	What is included in the concept?
		P	What items can the concept refer to?
	associations	R	What other words does this make us think of?
		P	What other words could we use instead of this one?
Use	grammatical functions	R	In what patterns does the word occur?
		P	In what patterns must we use this word?
	collocations	R	What words or types of words occur with this one?
		P	What words or types of words must we use with this one?
	constraints on use (register, frequency ...)	R	Where, when, and how often would we expect to meet this word?
		P	Where, when, and how often can we use this word?

Note: In column 3, R = receptive knowledge, P = productive knowledge.

Receptiv ordanvändning anses allmänt lättare än produktiv även om det inte är helt klart varför det förhåller sig så. Det kan emellertid, enligt Nation (2001), finnas många bidragande orsaker till detta. Vissa forskare hänvisar till det faktum att produktiv kunskap förutsätter mer precis kunskap om ordens såväl fonologiska, grafematiska som grammatiska egenskaper än receptiv kunskap. Man kan känna igen och identifiera ett ord i tal och skrift utan att för den skull kunna producera och använda det själv. En annan förklaring har att göra med att man normalt får mer övning i att använda orden receptivt än produktivt. Vissa ord kan också trots att de behärskas produktivt endast komma att användas receptivt genom att de av olika skäl inte kommer till aktiv användning. Ytterligare en annan förklaring kan ligga i att produktiv aktivering är mer krävande än receptiv, vilket har att göra med att kopplingen mellan ett ord i andraspråket och ett ord i förstaspråket (L2→L1=receptivt) på tidigare stadier är mer entydig och direkt än mellan ett ord i förstaspråket och ett ord i andraspråket (L1→L2=produktivt). Det beror på att det lexikala systemet i förstaspråket är betydigt mer utvecklat och att orden i förstaspråket ingår i komplexa nätverk som kan distrahera aktiveringen av orden i andraspråket. Flera studier visar också att receptiv inlärning av ord går fortare än produktiv liksom att testtagare får högre poäng på receptiva

än produktiva ordtest och att variationen mellan olika inlärares förmåga att lära nya ord är stor. Det kan också vara stor skillnad mellan en inlärares receptiva och produktiva förmåga att lära nya ord.

1.6 Mätning av ordförrådet i andraspråket

Erfarenheten visar att traditionella språktest inte på ett tillfredsställande sätt fångar upp det språk som utmärker skolans språkliga aktiviteter. Det har att göra med att sådana test ofta är inriktade på en mer allmän informell och vardaglig språkanvändning, och därför fungerar dåligt för mätning av språkfärdigheter med relevans för skolframgång. Testning och bedömning av skolrelaterad språkkunskap faller inom ramen för den testtradition som kallas LSP, d.v.s. *Language for Specific Purposes* som brukar särskiljas från testning av mer allmänna språkfärdigheter (Douglas 2000). Inom denna tradition betonas särskilt vikten av att de testuppgifter som används är representativa och relevanta inom det fält som testet gäller. Vidare understryker man interaktionen mellan språkkunskaper och fackkunskaper inom det område som testet gäller. Även om denna typ av test har utarbetats för en rad olika syften, sammanhang och språk har de nästan uteslutande utvecklats för att testa vuxna språkinlärares språkkunskaper för olika yrken och utbildningar. Språktest för bedömning av behörighet till universitetsstudier är exempel på test av akademiska språkfärdigheter som utvecklats inom denna tradition. I Sverige används t.ex. det s.k. TISUS-testet för detta ändamål vid olika universitet (se vidare http://www.nordiska.su.se/pub/jsp/polypoly.jsp?d=1668). Här testas ordkunskapen indirekt i olika läs-, hör-, tal- och skrivfärdighetsuppgifter men inte genom något speciellt ordtest. Den kartläggning av ordförrådet i läromedel för grundskolans senare år som presenteras i denna rapport har många tänkbara tillämpningar, inte minst när det gäller att utveckla diagnostiska modeller för mätning av ett skolrelaterat ordförråd. I det följande redovisas ett antal testformat som kan vara intressanta att utpröva med utgångspunkt från studiens kartläggning[2].

[2] I kapitel 4 i denna rapport redovisas ett antal genomförda utprövningar av ordtest som konstruerats utifrån studiens kartläggning.

1.6.1 Kvantitativa test

The Vocabulary Levels Test

Ett välkänt test för mätning av det akademiska ordförrådets storlek är *The Vocabulary Levels Test (VLT)*. Det har utarbetats av Paul Nation (1983, 1990) och använts i många olika sammanhang i och utanför Nya Zeeland. Det utvecklades ursprungligen i samband med en intensivkurs i engelska för internationella studenter som förberedelse för universitetsstudier. Testet har därefter använts för kvantitativa bedömningar av ordförrådet i diagnostiskt syfte också inom skolans värld i olika engelskspråkiga kontexter. I ett av de format som används ska sex ord matchas mot tre definitioner. Orden eller snarare definitionerna testas isolerade utan kontext som i nedanstående exempel:

1.	business	
2.	clock	___ part of a house
3.	horse	
4.	pencil	___ animal with four legs
5.	shoe	
6.	wall	___ something used for writing

VLT som också finns i versioner som mäter mer produktiva ordkunskaper (Laufer & Nation 1999) består av flera på varandra följande delar som innehåller ord på olika frekvensnivåer (se vidare kapitel 5 i denna rapport). Först testas ord bland de 2.000 vanligaste, sedan ord från frekvensbandet med de 2.001 – 3.000 vanligaste orden, sedan bland de 3.001 – 4.000 vanligaste orden o.s.v. I en (utvecklad) version av VLT som presenterats av Schmitt, Schmitt & Clapham (2001) består testet av ord från fem olika nivåer: de vanligaste 2000, 3000, 5000 orden + ord från den s.k. Academic Word List (AWL)[3] som återfinns bland de 5.001 – 10.000 vanligaste samt lågfrekventa ord utöver de 10.000 vanligaste som huvudsakligen representerar specifika ämnesområden. Ytterligare versioner av VLT finns tillgängliga på nätet bl.a. på http://www.er.uqam.ca/nobel/r21270/levels/ .

[3] Se avsnitt 1.3.1

Testet har även med framgång använts för att mäta ordförrådet i engelska som andraspråk hos skolelever i olika åldrar i Storbritannien (Cameron 2002). Resultaten visar att testet ger en god bild av olika elevgruppers receptiva ordförråd och viktig information om enskilda elever språkutveckling med stor relevans för forskning och för bedömning av behovet av pedagogiska insatser. Resultaten visar på signifikanta skillnader mellan första- och andraspråkselever. De senare uppvisar fortfarande efter tio år i brittisk skola luckor i fråga om det mest frekventa ordförrådet och har även problem med mindre frekventa ord, vilket naturligtvis får allvarliga implikationer för dessa elevers möjligheter att lyckas i skolan.

The Lexical Frequency Profile

Ett annat instrument som ursprungligen konstruerades av Laufer och Nation (1995) för mätning och analys av produktivt ordförråd i andraspråksskribenters texter är Lexical Frequency Profile (LFP). Detta verktyg har validerats i förhållande till andra lexikala test och uppvisat hög korrelation med etablerade mått på lexikal kunskap t.ex. det som uppnås genom VLT. LFP har använts för studier av ordförrådets utveckling samt för analys av helt andra typer av texter i olika syften sedan den gjorts tillgänglig på http://www.lextutor.ca/vp/ . Här kan vilken text som helst ”klistras in” varefter textens lexikala profil genereras på några sekunder. LFP visar den relativa andel ord en text innehåller från olika frekvensnivåer samt från den tidigare refererade Academic Word List (AWL). Beräkning av LFP görs med hjälp av datorprogrammet *VocabProfile* genom att textens ord matchas med frekvenslistor och fördelas på olika frekvensband. Andelen ord på olika frekvensnivåer läggs sedan till grund för en omvandling till procentsatser i förhållande till samtliga ord i texten. Nedan ges ett exempel på fördelning av ord efter frekvensband för en ”inklistrad” text.

As long as **issues** related to language and learning of ***bilingual*** students are not considered the ***responsibility*** of all teachers and **traditional ideologies** and *practices* are **maintained,** we cannot hope for any *improvement* as far as the **academic** success of ***bilingual*** students is concerned. Each new school subject ***entails*** a meeting with a new language and new *patterns* of language use characterized by **specific** and partly subject **specific** language. An important **aspect** of becoming a history ***physics*** or ***geography*** teacher is thus to become **aware** of the ***linguistic*** **dimensions** of the meaning **construction** in different *disciplines,* a *neglected* **aspect** in teacher *education* in ***Sweden*** and many other countries

1-1000 vanligaste orden
Ord bland 1001 – 2000 vanligase
Ord från Academic Word List
Övriga ord

1.6.2 Kvalitativa test

The Vocabulary Knowledge Scale

För mer kvalitativa bedömningar av ordförrådets djup i andraspråket finns relativt få utarbetade test. Ett sådant test är dock *The Vocabulary Knowledge Scale* (VKS) som utvecklats av Paribakht och Weshe (1997) för testning av ”icke-styrd” (incidental) ordinlärning. Testet bygger på två skalor; den ena för att elicitera svar och den andra för bedömning av svaren. För varje svar som testtagarna avger anger de också hur väl de känner till ordet på en skala indelad i fem kategorier från ”inte sett eller hört ordet” till ”jag kan använda ordet i en mening + exempel). Genom användning av en speciell bedömningsskala översätts sedan testtagarens svar till testpoäng.

Den första skalan som presenteras för testtagarna tillsammans med en lista av ord ser ut på följande sätt:

Tabell 1.6 VKS eliciteringsskala. (Paribahkt & Wesche 1997:180)

Self report categories	
I.	I don't remember having seen this word before.
II.	I have seen this word before but I don't know what it means.
III.	I have seen this word before, and I think I know what it means _____(synonym or translation).
IV.	I know this word. It means _____ (synonym or translation).
V.	I can use this word in a sentence: _____ (Write a sentence)

Den andra skalan som översätter testtagarens svar till ett testresultat ser ut på följande sätt:

Tabell 1.7 VKS poängberäkningskategorier. (Paribahkt & Wesche 1997:180)

Self report categories	**Possible scores**	**Meaning of scores**
I	1	This word is not familiar at all.
II	2	The word is familiar but its meaning is not known.
III	3	A correct synonym or translation is given.
IV	4	The word is used with semantic appropriateness and grammatical accuracy in a sentence.

Associationstest

En annan metod att mäta mer kvalitativa aspekter av ordförrådet i andraspråket är genom s.k. *associationstest* (Meara 1983). Här ombeds testtagarna säga/skriva det eller de ord de först kommer att tänka på när de hör/ser ett visst ord. I vissa test uppmanas testtagarna i stället välja utifrån en given lista av tänkbara associationer. Ordassociationer utgör länkar som på något sätt relaterar ord i språkanvändarens mentala lexikon. Jämförelser mellan de associationer som ges av

andraspråksinlärares och infödda språkanvändare vid denna typ av test har visat att andraspråksinlärare trots sin mer begränsade vokabulär tenderar att ge betydigt mer varierade svar än infödda språkbrukare som i allmänhet avger mer samlade och stereotypa svar. Detta anses tyda på ett ordförråd med mer etablerade och befästa lexikala nätverk. En annan viktig skillnad är att andraspråkinlärare speciellt i tidigare stadier tenderar att göra s.k. klangassociationer, d.v.s. välja ord som ljudmässigt snarare än betydelsemässigt kan relateras till stimulusordet, vilket är betydligt mer ovanligt hos infödda språkbrukare (jfr Namei 2002).

Tine Greidanus & Lydius Nienhuis (2001) redovisar en studie av ordförrådsutveckling hos avancerade inlärare av franska med engelska och nederländska som modersmål där associtionstest användes. De 50 stimulusorden hämtades här från de femtusen vanligaste orden uppdelade i fem frekvensband. Stimulusorden bestod av såväl konkreta som abstrakta substantiv, verb och adjektiv och associationsorden var antingen hämtade från samma frekvensband som stimulusorden eller från frekvensband med mer frekventa ord. Testet som hade som främsta syfte att mäta ordförrådsutveckling prövade också effekten av olika typer av distraktorer, nämligen sådana som var semantiskt relaterade till stimulusorden och sådana som inte var det. I de flesta tidigare test har distraktorerna i allmänhet valts bland icke semantiskt relaterade ord något som man i detta sammanhang bedömde som mindre lämpligt med tanke på testtagarnas avancerade nivå. För varje stimulusord kunde testtagarna välja mellan tre associationsord med följande semantiska relationer med stimulusorden + tre distraktorer:

- **Paradigmatiska**: hyperonymer, hyponymer, synonymer och antonymer
  Exempel: hund → husdjur, tax, jycke, katt
- **Syntagmatiska**: kollokationer
  Exempel: hund → skäller, morrar, gläfser, renrasig
- **Analytiska**: karakteristika som i ordboksdefinitioner
  Exempel: hund → fyrbent, svansförsedd

Följande två exempel är hämtade från de två versionerna av testet (semantiskt relaterade resp. icke-relaterade ord):

rive **(strand)**	*rive* **(strand)**
☐ *artificiel* (artificiell)	☐ *bord* (sida)
☐ *bord* (strand)	☐ *fleuve* (flod)
☐ *côté* (sida)	☐ *gauche* (vänstra)
☐ *fleuve* (flod)	☐ *pacquet* (paket)
☐ *gauche* (vänstra)	☐ *prudent* (noggrann)
☐ *vague* (våg)	☐ *tombe* (grav)

De associationsord som testades i båda versionerna var *bord* (paradigmatiskt relaterat), *gauche* (syntagmatiskt relaterat) och *flueve* (analytiskt relaterat). Det vänstra testet innehåller tre semantiskt relaterade distraktorer (*artificiel*, *côté* och *vague*); det högra testet innehåller tre semaniskt icke-relaterade distraktorer (*parcel, prudent* och *tombe*).

Testen diskriminerade väl mellan testtagare på olika nivåer och de semantiskt relaterade distraktorerna befanns mer lämpade att fånga kvalitativa skillnader i ordkunskap hos avancerade inlärare än de icke-semantiskt relaterade eftersom de i högre grad ”lockade” testtagarna att välja dem. Som väntat med tanke på målgruppen föredrogs paradigmatiska associationsord. Dessutom valdes analytiska associationsord i högre utsträckning än syntagmatiska. Det fanns också ett klart samband mellan frekvens och kvalitet på testtagarnas ordkunskap. Ju vanligare stimulusordet var desto bättre var testtagarnas kunskap om ordet, vilket innebär att ordförrådets djup inte utvecklas i samma takt som ordförrådets bredd. Kunskapen om tidigt inlärda vanliga ord är djupare än kunskapen om senare inlärda mer ovanliga ord.

Depth of Vocabulary Test

Andra aspekter som fokuseras i test inriktade på kvalitativa aspekter av lexikal kompetens är ordbetydelse (d.v.s. främst synonymi och polysemi) och kollokationer. Qian & Schedl (2004) har utvecklat ett *Depth of Vocabulary Knowledge Test* (DVK) för engelska som andraspråk baserat på 40 ”items” bestående av ett stimulusord som är ett adjektiv

och två boxar innehållande fyra ord vardera. Bland de fyra orden i den första boxen (kallad DVK-betydelse) kan ett till tre ord vara en synonym till eller i vissa kontexter utbytbart med stimulusordet medan ett till tre av de fyra orden i den andra boxen (kallad DVK-kollokation) kan ingå i en kollokation med stimulusordet[4]. Varje item har fyra korrekta svarsalternativ som emellertid kan vara olika distribuerade i de två boxarna. Ett exempel på hur formatet utformas är:

Powerful

(A) potent (B) definite (C) influential (D) supportive	(E) position (F) engine (G) repetition (H) price

Answer: (A) (B) (C) (D) (E) (F) (G) (H)

Vid bedömningen ger varje korrekt svarsalternativ en poäng och maximalt antal poäng på testet är alltså 4 x 40 = 160 poäng. Testtagaren får 35 minuter till sitt förfogande för att utföra testet. Qian & Schedl har visat att testresultaten på DVK-testet korrelerar väl med resultaten på det s.k. TOEFL-VOC-testet som ingår i det välkända TOEFL-testet som tillämpas för mätning av språkfärdigheter i engelska som andraspråk över hela världen liksom med testresultaten på läsförståelsedelen TOEFL-RBC (Reading for Basic Comprehension) i samma test. Bortsett från vissa svårigheter med testkonstruktionen som bl.a. har att göra med att ge testtagarna entydiga och begripliga instruktioner samt med att hitta okontroversiella korrekta svarsalternativ i kollokationsboxarna menar Qian & Schedl att testet kan ha stor potential att så småningom ersätta det mer traditionella och endimensionella TOEFL-VOC-testet som bygger på ett flervalsformat med ”items” av följande typ:

Tung oil is a *powerful* drying agent in varnishes and paints.

(B) pure
(C) potent
(D) poisonous
(E) permanent

[4] För en diskussion av flerordsenheter se kapitel 2 i denna rapport.

1.7 Evidensbaserad diagnostisering av skolrelaterade språkfärdigheter

I samband med en nyligen genomförd skolreform i USA refererad till som *No children left behind ACT* (2001), har bedömning av språkfärdigheter för elever med andra modersmål blivit en het fråga eftersom reformen kräver att skolorna måste visa att elever med engelska som andraspråk gör mätbara framsteg i sin engelskutveckling varje år. Genom olika studier har man dock funnit att befintliga språktest är inadekvata i detta sammanhang eftersom de ofta inte mäter sådana språkfärdigheter som är avgörande för elevernas skolframgång utan snarare är inriktade på en mer informell vardaglig språkanvändning där det språk som utmärker hela den vidd av språkliga praktiker i skolan inte fångas upp. Mot bakgrund av detta pågår nu ett arbete vid Centre for Evaluation, Standards and Student Testing (CRESST) vid UCLA (Butler, Lord, Stevens, Borrego & Bailey, 2004) som syftar till att framställa bedömningsinstrument som mer direkt inriktas på centrala aspekter på skolspråket, d.v.s. på det man kallar *Academic English.*

Utgångspunkten för arbetet är att man, för att kunna beskriva och operationalisera skolspråket som teoretisk konstrukt, systematiskt måste dokumentera det språk som eleverna förväntas kunna ta del av och själva producera för att lyckas i skolan. Relevant empiriskt underlag för en sådan beskrivning av det ”akademiska” skolspråket utgörs t.ex. av målbeskrivningar i styrdokument och nationella prov för olika ämnen, läromedel samt klassrumsinspelningar. Butler et al. beskriver utvecklingen av valida och reliabla testinstrument som en i högsta grad strukturerad process i flera led som tar sin utgångspunkt i en behovsanalys och en väl definierad teoretisk modell över den kompetens och de färdigheter som ska testas. Denna teoretiska ram läggs sedan till grund för utvecklingen av de testspecifikationer och prototypuppgifter som lärare och testkonstruktörer kan använda för att själva utveckla relevanta specifikationer och uppgifter som svarar mot specifika behov av operationalisering av akademisk engelska i olika utbildningskontexter.

Tabell 1.8 Innehåll i uppgiftsspecifikationer (Butler et al.: 2004:62).

Reading skills
(1) Identify main idea (2) Locate supporting detail
Language functions (with embedded grammatical features)
(1) Comparison/contrast (a) Adverbial comparison (b) Comparative adjective forms (c) Logical connectors (2) Description (a) Logical connectors (b) Nominal structures (c) Passive voice (d) Prepositions (e) Simple present tense (f) Subordinate clauses (3) Explanation (a) Logical connectors
Vocabulary
(1) Identify meaning in context (2) Draw meaning from embedded definition(s)

I ett första led har tonvikten lagts vid läsning men det fortsatta arbetet kommer även att inriktas på skriv- och höruppgifter och uppgifter inriktade på mer integrerade färdigheter. Med utgångspunkt från autentiska texter i matematik och NO-ämnen formuleras uppgifter med fokus på sådana språkliga drag som kännetecknar språkbruket i dessa ämnen. Vanliga uppgiftstyper som eleverna förväntas kunna hantera här består i att komplettera uppgifter i grafiska figurer, att utforma listor, matcha uppgifter i olika kolumner, fylla i flervals- och kortsvarsfrågor. Val av textmaterial görs utifrån detaljerade kriterier för textkällor, textlängd, språkfunktioner och ämnesinnehåll. Som framgår av tabell 1.8 har man för arbetet med specifikationer och uppgiftsprototyper i detta första led utgått från tre fokusområden: (1) läsfärdigheter, (2) språkfunktioner (och med dessa förknippade grammatiska drag) och (3) vokabulär.

Genom utveckling av evidensbaserade och kvalificerade modeller för diagnostisering av skolrelaterade språkfärdigheter kan språkliga behov hos elever som inte fullt ut behärskar undervisningsspråket kartläggas på ett relevant, effektivt och systematiskt sätt som underlag för olika pedagogiska insatser. Här utgör det arbete som genomförs vid CRESST av Butler och hennes kollegor en mycket intressant modell för framtida svenska insatser.

I detta inledande kapitel har vi redovisat forskning som i mycket vid mening kan ses som en bakgrund till arbetet i OrdiL-projektet genom att den på olika sätt anknyter till projektets syfte, nämligen att öka kunskapen om lärandets språkliga dimensioner ur ett andraspråksperspektiv. Även om tyngdpunken i det här aktuella OrdiL-projektet ligger på lexikaliska aspekter på läromedelsspråket är tanken bakom uppbyggnaden av OrdiL-korpusen att den framöver ska kunna användas för språkliga analyser av läromedelsspråket av många olika slag och för en rad olika forsknings- och undervisningsändamål. Här kan de studier som aktualiserats i denna forskningsöversikt bidra med intressanta infallsvinklar. Det är därför vår förhoppning att denna projektrapport ska ge inspiration till flera intressanta tillämpningar av korpusen i framtiden.

Referenser

Allén, S. 1970. *Nusvensk frekvensordbok baserad på tidningstext*. Stockholm: Almqvist & Wiksell.

Allén, S. 1972. *Tiotusen i topp: ordfrekvenser i tidningstext*. Stockholm: Almqvist & Wiksell.

Allén, S., Berg, S., Järborg, J., Löfström, J., Ralph, B. & Sjögren, C. 1980. *Nusvensk frekvensordbok 4*. Ordled, betydelser. Stockholm: Almqvist & Wiksell.

Bachman, L. & Palmer, A. 1996. *Language testing in practice*. Oxford: Oxford University Press.

Biber, 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, D., Conrad, S., & Reppen, R. 1998. *Corpus linguistics: Investigating structure and use*. Cambridge: Cambridge University Press.

Björnson, C.H. 1968. *Läsbarhet*. Stockholm: Liber

Butler, F., Lord, C., Stevens, R., Borrego, M. & Bailey, A. 2004. *An approach to operationalizing academic language for language development purposes: Evidence from fifth-grad science and math*. CSE report 626. CRESST, UCLA.

Cameron, L. 2002. Measuring vocabulary size in English as an additional language. *Language Teaching Research* 6 (2), 145-173.

Carlo, M., August, D., McLaughlin, B., Snow, C., Dressler, C., Lippman, D., Lively, T. & White, C. 2004. Closing the gap: Addressing the vocabulary needs of English-language learners in bilingual and mainstream classrooms. *Reading Research Quarterly,* 39 (2), 188-215.

Carroll, J., Davies, P. & Richman, B. 1971. *The American heritage word frequency book*. New York: Houghton Mifflin, Boston American Heritage.

Carver, R. 1994. Percentage of unknown vocabulary words in a text as a function of the relative difficulty of the text: implications for instruction. *Journal of Reading Behavior,* 26, 413-437.

Corson, D. 1985. *The Lexical Bar*. Oxford: Pergamon Press.

Corson, D. 1995. *Using English words,* Dordrecht: Kluwer Academic Publishers.

Coxhead, A. 1998. *An academic word list*. Wellington, New Zealand: Victoria University of Wellington.

Coxhead, A.J. 2000. A new academic word list. TESOL *Quarterly,* 34:2.

Danielson, S. 1975. Läroboksspråk. *En undersökning av språket i vissa läroböcker för högstadium och gymnasium.* Umeå: Acta Universitatis Umensis.

Donley, K. & Reppen, R. 2001. Using corpus tools to highlight academic vocabulary i SCILT. *TESOL Journal,* Vol. 10: 2-3.

Douglas, D. 2000. *Assessing languages for specific purposes.* Cambridge: Cambridge University Press.

Edling, A. 2004. Abstraktion kan spränga gränser. *Språkvård 3*, 2004.

Ekvall, U. 1995. Läroboken - begriplig och intressant? I Siv Strömquist (red.) *Läroboksspråk.* Uppsala: Hallgren och Fallgren.

Enström I. 2004. Ordförråd och ordinlärning med särkilt fokus på avancerade inlärare. I Kenneth Hyltenstam och Inger Lindberg (red.) *Svenska som andraspråk i forskning, undervisning och samhälle.* Lund: Studentlitteratur.

Flesch, R. 1949. *The Art of Readable Writing.* New York: Harper & Row.

Fredriksson, U. & Taube, K. 2001. Läsning bland elever med invandrarbakgrund. *En undersökning av läsförmåga och bakgrundsfaktorer hos elever i årskurs 3 i Stockholm.* Stockholms universitet, Institutionen för internationell pedagogik.

Frick, N. & Malmström, S. 1976. *Språkklyftan. Hur 700 ord förstås och missförstås.* Kristianstad: Tidens förlag.

Gimbel, J. 1995. Bakke og udale. *Sprogforum 3.*

Golden, A. 2005. *Å gripe poenget. Forståelse av metaforiske uttrykk fra lærebøker i samfunnskunnskap hos minoritetselever i ungdomsskolen.* Acta Humaniora 227. Oslo: UniPub Forlag.

Golden, A. & Hvenekilde, A. 1983. *Rapport fra prosjektet Læbokspråk. Sentret för språkpedagogikk,* Universitetet i Oslo.

Greidanus, T. & Nienhuis, L. 2001. Testing the quality of word knowledge in a second language by means of word association: type of distractors and types of associations. *Modern Language Journal* 85:567-77.

Halliday, M. 1994. *An introduction to functional grammar.* 2 utg. London. Melbourne & Auckland: Edward Arnold.

Hene, B. 2004. Adjektivs metaforiska betydelse – utlandsadopterade och svenska barns tolkningar. I K. Hyltenstam & I. Lindberg (red.) *Svenska som andraspråk i forskning, undervisning och samhälle.* Lund: Studentlitteratur.

Henriksen, B. 1999. Three dimensions of vocabulary development. *Studies in Second Language Acquisition* 21, 303-17.

Huckin, T., Haynes, M. & Coady, J. 1995. *Second language reading and vocabulary learning.* Norwood, NJ: Ablex.

Hvenekilde, A. 1986. Nærblikk på o-fags-tekster. *Norsklæreren 3.*

Hvenekilde, A. (red.) 1991. *Matte på ett språk vi förstår.* Stockholm: Scriptor. Internationellt perspektiv.

Hägerfelth, G. 2004. Språkpraktiker i naturkunskap i två mångkulturella gymnasieklassrum. En studie av läroprocesser bland elever med olika förstaspråk. *Malmö Studies in Educational sciences.* No. 11.

Kulbrandstad, L. 1996. Lesing på et andrespråk – *En studie av fire innvandrerungdommers lesing av lærebokstekster på norsk.* Oslo: Scandinavian University Press.

Kulbrandstad, L. 2003. *Lesing i utvikling. Teoretiske og didaktiske perspektiver.* Bergen: Fagbokforlaget.

Laufer, B. & Nation, P. 1995. Vocabulary size and use: lexical richness in L2 written production. *Applied Linguistics,* 16, 307-322.

Laufer, B. & Nation, P. 1999. A vocabulary size test of controlled productive ability. *Language Testing* 16, 36-55.

Laursen, H. 2004. *Den språglige dimension i naturfagsundervisningen – fokus på det flersprågede klasserum.* København: Københavns kommune, CVU.

Lexin, *Språklexikon för invandrare.* Stockholm: Natur & Kultur. http://www-lexikon.nada.kth.se/skolverket/lexin.shtml 2006-06-30.

Lindberg, I. 2004. Nationella provet i svenska för invandrare. Utgångspunkter för provkonstruktionen. http://www.biling.su.se/~sfi/ingers-text.pdf 2006-06-30.

Lively, B. & Pressey, S. 1923. A Method for Measuring the 'Vocabulary Burden' of textbooks", *Educational Administration and Supervision,* 9: 389-398.

Meara, P. 1983. *Word associations in a foreign language.* A report of the Birkbeck Vocabulary Project. Nottingham Linguistic Circular, 11, 29-38.

Meara, P. 1990. A note on passive vocabulary. *Second Language Research,* 6:150-154.

Melin, L. 1992: *Textbindning och läsbarhet.* I Svenskans beskrivning 19. Lund.

Melin, L. 1995. Grafisk pyttipanna. I Siv Strömquist (red.) *Läroboksspråk*. Uppsala: Hallgren och Fallgren.

Myndigheten för Skolutveckling 2004. *Kartläggning av svenska som andraspråk.*

Namei, S. 2002. *The bilingual lexicon from a developmental perspective.* Centrum för tvåspråkighetsforskning. Stockholms universitet.

Nation, P. 1983. *Testing and teaching vocabulary.* Guidelines 5, 12-25.

Nation, P. 2001. *Learning vocabulary in another language.* Cambridge: Cambridge University Press.

Nation, P. 1990. *Teaching and learning vocabulary.* New York: Heinle and Heinle.

Olvegård, L. 2006. *Vad händer i skolans historieböcker? En undersökning om innehåll och språkliga strukturer i lärobokstexter i historia*. Specialarbete i svenska som andraspråk 61-80 p, Institutionen för svenska språket, Göteborgs universitet.

Paribakht, T. & Wesche, M. 1997. *Vocabulary enhancement activities and reading for meaning in second language vocabulary acquisition.* I Coady & Huckin (eds.) Second language vocabulary. Cambridge: Cambridge University Press.

Parszyk, I-M. 1999. En skola för andra. Minoritetselevers upplevelser av arbets- och livsvillkor i grundskolan. *Studies i educational sciences* 17. Stockholm: HLS Publications.

Qian, D. 2002. Investigating he relationship between vocabulary knowledge and academic reading performance: an assessment perspective. *Language Learning* 52:513-36

Qian, D. & Schedl, M. 2004. Evaluation of an in-depth vocabulary knowledge measure for assessing reading performance. *Language Testing* 21 (1), 28-52. 2002

Read, J. 2000. *Assessing vocabulary.* Cambridge: Cambridge University Press.

Reichenberg, M. 2000. *Röst och kausalitet i lärobokstexter. En studie av elevers förståelse av olika textversioner.* Acta Universitatis Gotoburgensis.

Richards, J. 1976. The role of vocabulary teaching. *TESOL Quarterly* 10, 77-89.

Rothery, J. 1996. Making changes: developing an educational linguistics. I Ruqaiya Hasan & Geoff Williams (red.) *Language in Society*. London: Longman.

Rönnberg, I. & Rönnberg, L. 2001. *Minoritetselever och matematikutbildning - en litteraturöversikt*. Stockholm: Skolverket.

Saville-Troike, M. 1984. What really matters in second language learning for academic achievement? *TESOL Quarterly* 18:2.

Schmitt, N., Schmitt, D. & Clapham, C. 2001. Developing and exploring the behaviour of two versions of the Vocabulary Levels Test, *Language Testing* 18 (1).

Skolverket 2006. *Läroplan för det obligatoriska skolväsendet, förskoleklassen och fritidshemmet*.

SOU 2004:97. *Att lyfta matematiken*. Betänkande av matematikdelegationen. Utbildningsdepartementet.

Stroud, C. 1979. Kontrastiv lexikologi. I: Hyltenstam, K. (utg.), *Svenska i invandrarperspektiv. Kontrastiv analys och språktypologi*. Lund: Liber läromedel.

Swanborn, M. & De Glopper, K. 1999. Incidental word learning while reading: A meta-analysis. *Review of Educational Research*, 69:261-286.

Thomas & Collier, 1997. *School effectiveness for language minority students*. Washington, DC: National Clearinghouse for Bilingual Education.

Thorndyke, E. 1921. *The Teacher's Word Book*. Teachers College, Columbia University. First Edition. University Press.

Torvatn, A-C. 2002. *Tekststrukturens innvirkning på leseforståelsen - en studie av fire læreboktekster for ungdomstrinnet og sju elevers lesing av dem*. Institutt for språk og kommunikasjonsstudier, NTNU.

Verhallen, M. & Schoonen, R. 1993. Lexical knowledge of monolingual and bilingual children. *Applied Linguistics,* 13.

Verhoeven, I. & Vermeer, A. 1985. Ethnic group differences in children's oral proficiency in Dutch. I G. Extra & T. Vallen (eds.) *Ethnic minorities and Dutch as a second language*. Dordrecht/ Holland: Foris Publications.

Vermeer, A. 2001. Breadth and depth of vocabulary in relation to L1/ L2 acquisition and frequency of input. *Applied Psycholinguistics*, 22:2.

Viberg, Å. 1993. Andraspråksinlärning i olika åldrar. I E. Cerú (red.) *Svenska som andraspråk. Lärarbok 2*. Stockholm: Natur & Kultur.

Viberg, Å. 2004. Lexikal utveckling i ett andraspråk. I K. Hyltenstam & I. Lindberg (red.) *Svenska som andraspråk i forskning, undervisning och samhälle*. Lund: Studentlitteratur.

Vygotsky, L. 1978. *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

Vygotskij, L. 1999. *Tänkande och språk*. Göteborg: Daidalos.

West, M. 1953. *A general service list of English words*. London: Longman, Green & Co.

Westman, M. 1974. *Bruksprosa*. Lund: Liber Läromedel.

Zareva, A., Schwanenflugel, P. & Nikolova, Y. 2005. Relationship between lexical competence and language proficiency. *Studies in Second Language Acquisition*, 27, 567-595.

Webblänkar

http://www.nordiska.su.se/pub/jsp/polypoly.jsp?d=1668/ 2007-02-19
http://www.er.uqam.ca/nobel/r21270/levels/ 2007-02-19
http://www.lextutor.ca/vp/ 2007-02-19

2. Om ord och ordkunskap

Jerker Järborg

I projektet Ord i läromedel (OrdiL) studeras ord och ordförråd eller vokabulärer. Både ”ord” och ”vokabulärer”/”ordförråd” visar sig emellertid vara ganska komplicerade begrepp, vilka fordrar särskilda teoretiska preciseringar för att kunna studeras vetenskapligt. För en närmare utredning av begreppet ’ord’ hänvisas till nästa avsnitt; tills vidare används ”ord” som en bekväm beteckning dels för de enheter man möter i texter, dels för de enheter som brukar förtecknas i ordböcker och som kan antagas ingå i olika individers ordförråd, vilka i sin tur utgör en viktig del av deras språkkunskaper. Denna sistnämnda aspekt får vara utgångspunkt för den fortsatta diskussionen.

Projektets syfte är, som framgått av kapitel 1, i första hand att skapa ett underlag för bedömning av svårighetsgraden av ordförrådet i läromedel, särskilt med tanke på de speciella svårigheterna för skolelever med ett annat modersmål än svenska. I andra hand är avsikten att visa hur detta underlag kan användas för att t.ex. testa ordkunskap eller underlätta för lärare och läroboksförfattare. Innan man kan börja analysera ett faktiskt ordförråd i ett textmaterial i dessa avseenden måste man dock utreda vad man vet om ordkunskap och ordförståelse hos enskilda språkbrukare. I annat fall är det lätt hänt att man gör naiva antaganden om ”svårigheter” som inte existerar – eller tvärtom, att man missar verkliga svårigheter som

inte är omedelbart synliga vid en ytlig betraktelse. T.ex. kan det finnas det föreställningar om att alla lågfrekventa ord i ett större textmaterial måste vara föga förtrogna eller ”svåra” – något som *inte* gäller generellt.

2.1 Det mentala lexikonet

När ”språket” som abstrakt helhet diskuteras talas ofta om variation, som mellan tal- och skriftspråk, mellan samtalsspråk kring kaffebordet, yrkesjargong och språket i klassrummet, mellan mans- och kvinnospråk, mellan genrer som lagspråk, skönlitterärt språk, tidningsspråk etc. och inte minst mellan den för oss viktigaste distinktionen: allmänspråk och fackspråk. I själva verket finns inga skarpa gränser mellan sådana språktyper och de skillnader som finns är vanligen små och kan ligga både i grammatiken och i ordförrådet. Även för fackspråk gäller att den absolut största delen av ordförrådet utgörs av allmänspråkliga ord, dock ibland med en annan frekvens än i allmänspråket. Dessutom är många facktermer i princip specialiseringar av allmänspråkliga ord, som vi kommer att se i det följande. Det finns alltså ingen teoretisk anledning att tänka sig det mentala lexikonet uppdelat i t.ex. en ”allmän” och ett antal ”fackspråkliga” delar. Däremot kommer naturligtvis de rena facktermerna normalt att tillhöra de ”yttre” lexikala skikten, vilka inlärs sent och till en början är föga förtrogna, enligt nedanstående beskrivning.

Det mentala lexikonet hos en ”genomsnittlig” språkbrukare växer under individens livstid och särskilt under skoltiden, så som beskrivs i kapitel 1. Tillväxten sker dels genom att nya ord inlärs, dels genom att kunskapen om varje ord ökar, dels genom att olika samband mellan ord i lexikonet tillkommer. Den första typen av tillväxt behöver knappast exemplifieras. Den andra typen kan innebära bl.a. att man lär sig sådant som att ett ord som *måla* kan ha olika betydelser (hantverk vs konst), att ett ord som *kraft* har en specialiserad betydelse inom fysiken eller att många ord kan ha överförda betydelser, som *djupdykning* ≈ ’inträngande undersökning’. Särskilt viktigt för förståelsen är

att lära sig de sammanhang eller s.k. semantiska kontexter som typiskt förbinds med vissa ord eller ordbetydelser: den ”konstnärliga” betydelsen av *måla* förbinds bl.a. med ett motiv (*måla av ngt/ngn*) och den som utför handlingen antages normalt vara just en konstnär. Den tredje typen kan innebära att man lär sig sådant som att en *häst* är ett *hovdjur* eller *riddjur*, att *hästar* kan vara *varmblod*, *ardenner* m.m., att man *rider på* eller *kör* häst och att man då använder bl.a. *sadel* och *tyglar* eller *sele* och *tömmar*. Man skulle kunna säga att lexikonet på detta sätt kommer att utveckla små nätverk av ord, i detta fall ett nätverk kring ordet *häst*. Man skall inte heller glömma den enklare formen, som kan kallas ordbildningsnätverk: till *måla* bl.a. *målning, målare, vitmålad, målarfärg* etc. Lexikonet är således inte enbart en lista av enheter. Det är därför inte tillräckligt att jämföra enbart storleken på språkbrukarens ordförråd; man borde också undersöka hur djupgående ordkunskaperna är, vilket är görbart men ofta ganska komplicerat. Vad själva storleken beträffar kan man lugnt antaga att en vuxen språkbrukare har ett mentalt lexikon för sitt modersmål på många tiotusentals ord, något som bekräftas av undersökningar, se kapitel 1. Undersökningarna tycks också visa att andraspråkslexikonet för ungdomar med annat modersmål generellt sett är betydligt mindre än ett modersmålslexikon men tillväxer i ungefär samma takt. Huruvida kunskaperna om andraspråkets ord är lika djupgående är väl knappast tillräckligt undersökt. Det finns tecken som tyder på dels att vissa förtrogna ord, se nedan, överanvänds, dels att modersmålets lexikon ibland kan styra användningen av ungefär likabetydande andraspråksord. Här fordras dock ytterligare forskning.

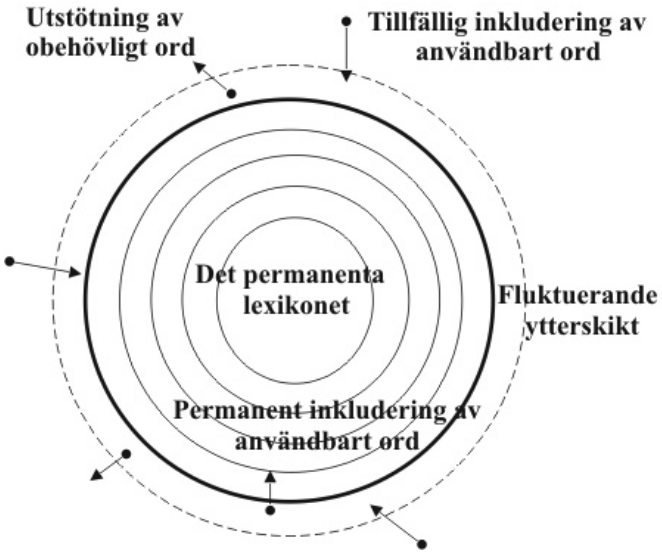
Det är ganska självklart att språkbrukaren bör ha olika grad av kunskap om orden i sitt mentala lexikon och att man därför kan tala om att vissa ord är mer förtrogna eller mer semantiskt centrala; andra mindre kända eller mer semantiskt perifera. Den mest kända uppdelningen är den som har brukat göras inom språkpedagogiken mellan den produktiva och den receptiva vokabulären, tidigare: aktiv respektive passiv vokabulär. Som beteckningarna anger består den enbart receptiva vokabulären av ord som en språkbrukare ”kan” i någon mening, t.ex. genom att ungefärligen känna till betydelsen, utan att kunskaperna dock är tillräckliga för att han/hon skall kunna använda ordet produktivt i egen språkproduktion. I enlighet med föregående

resonemang lär sig språkbrukaren normalt allt mer om orden i sin receptiva vokabulär, varvid de ”innersta” småningom övergår till den produktiva vokabulären. Samtidigt växer givetvis den receptiva vokabulären genom tillägg av nya ord. Det är dock troligt att s.k. tillfälliga sammansättningar, *ordförrådsundersökningsresultat* o.d., aldrig egentligen kommer att ingå i det mentala lexikonet utan i stället kan bildas och tolkas vid behov. Således måste förmågan att bilda och tolka sådana sammansättningar också räknas in i den lexikala kunskapen.

Det är naturligtvis mycket svårt att närmare bestämma hur det mentala lexikonet är uppbyggt och vilka ord som är mest centrala, särskilt som de semantiskt centrala orden inte nödvändigtvis är de mest förtrogna. Följande Rosch (1977) kan man gissa att de s.k. basnivåbegreppen är de mest förtrogna inom ett givet s.k. semantiskt fält. Sålunda skulle ord som *häst* eller *skåp* höra till basnivån medan *hovdjur* respektive *möbel* skulle vara mer övergripande och alltså mindre förtrogna och *ardenner* respektive *kommod* givetvis representerar en finare indelning och även de bör vara mindre förtrogna. Om man betraktar sviten *förflytta sig*, *gå*, *lunka* kan man se att det förtrogna verbet *gå* är mindre användbart för att beskriva liknande verbbetydelser (*krypa*, *åla*) än det övergripande *förflytta sig*; här har vi alltså ett exempel på en skillnad mellan förtrogenhet och semantisk centralitet.

Figur 2.1 Det mentala lexikonet i utveckling.

Det mentala lexikonets struktur under utvecklingen skulle kunna grovt illustreras som i figur 2.1. Den innersta cirkeln får här motsvara antingen de mest förtrogna eller de semantiskt centrala orden. De följande cirklarna ”utåt” motsvarar då ord om vilka språkbrukaren har successivt mindre kunskap. Gränsen mellan den inre, produktiva vokabulären och den receptiva vokabulären markeras med en kraftigare cirkellinje. Det närmast liggande vokabulärskiktet representerar de bäst kända orden i den receptiva vokabulären och kommer under den fortgående inlärningen att integreras i den produktiva vokabulären. Därigenom flyttas alltså gränsen ”utåt”, varefter processen kan tänkas upprepas.



Figur 2.2 Tillfälliga ord i lexikonet.

Vid en given tidpunkt kan man tänka sig det mentala lexikonet illustreras med en variant, nämligen figur 2.2. Här gör vi ingen skillnad på den produktiva och den receptiva vokabulären utan koncentrerar oss på det yttersta skiktet, mellan den kraftiga och den streckade cirkel-

linjen. Detta ytterskikt får representera de tillfälliga sammansättningar och avledningar med förutsägbar betydelse som språkbrukaren ständigt bildar och tolkar. Det är knappast troligt att sådana ord, som *ordförrådsundersökningsresultat* ovan, lagras permanent i det mentala lexikonet men inte heller troligt att de måste nybildas eller nytolkas varje gång de används i en given situation, t.ex. medan man läser denna text. Man får därför tänka sig att det mentala lexikonet alltid kompletteras av ett fluktuerande ytterskikt som i figuren, bestående av tillfälliga sammansättningar, etc., vilka endast lagras så länge språkbrukaren har direkt behov av dem. Innehållet i detta skikt varierar alltså efter situationen.

Ytterligare aspekter på ordkunskap kommer att tas upp i samband med utredningen av ordbegreppet i avsnitt 2.2 och framför allt i ett särskilt avsnitt 2.3 om lexikal kompetens. Man kan emellertid redan nu konstatera att otillräckligt utvecklade lexikala kunskaper kan ta sig många former och skulle kunna leda till många olika typer av bristande förståelse vid textläsning. Utmaningen är att försöka finna storskaliga metoder för att i någon mån identifiera de ord i ett textmaterial som kan antagas vara svårtolkade enligt ovanstående resonemang, t.ex. på grund av att de är flertydiga, perifera i lexikonet eller har en komplicerad semantik. Därvid får man försöka approximera vissa kvalitativa begrepp som 'förtrogenhet' med kvantitativa, som frekvens och spridning. För identifiering av ord som kan skapa mer utpräglade semantiska svårigheter måste dock de statistiska metoderna kompletteras genom insatser av mänsklig expertis. Projektets principmetoder diskuteras i avsnitt 2.4, för de tekniska detaljerna hänvisas till kapitel 3.

2.2 Ordbegreppet

För skriftspråk gäller att en teckenföljd som betraktas som *ett* ord konventionellt avgränsas med mellanrum eller blanktecken (spatium), ibland även med radskifte eller dylikt. Däremot gäller inte motsatsen: att alla teckenföljder som avgränsas på detta sätt också bör betraktas som riktiga ord, vilka skulle kunna återfinnas i en ordbok. En stor och viktig grupp av undantag utgörs av olika typer av egennamn t.ex.

personnamn, geografiska namn, namn på företag och produkter, titlar på filmer och böcker m.m. Namn återfinns, som bekant, vanligen inte i ordböcker på grund av att kunskap om namn brukar anses ligga utanför de språkliga kunskaperna och i stället tillhöra den s.k. omvärldskunskapen – den encyklopediska kunskapen. I en undersökning av ordförrådet i en text bör alltså egennamnen på lämpligt sätt undantas och eventuellt särbehandlas. Andra teckenföljder som bör undantas är sifferuttryck och förkortningar, se vidare nedan.

Även om de avgränsade teckenföljderna representerar ”riktiga” ord är det viktigt att hålla i minnet att de ord som påträffas i texter i allmänhet uppträder i någon böjningsform medan orden i ordböcker och ordlistor förtecknas i sin grundform. Därför kan uppgifter om antalet olika ord i en text inte utan vidare jämföras med t.ex. antalet ord i en ordbok. Även om grundformerna skulle ha kunnat identifieras kan de fortfarande vara flertydiga: de kan tillhöra olika ordklasser, som *kratta* (substantiv eller verb) och de kan, inom samma ordklass, ha olika betydelser, som *by* ’litet jordbrukarsamhälle’ vs ’vindstöt’. Slutligen förekommer vissa ord i mer eller mindre fasta förbindelser, s.k. kollokationer, i vilka de inte kan tolkas isolerade utan endast tillsammans med övriga ord i förbindelsen som tvåordsverbet *äga rum*, där varken *äga* eller *rum* har någon av sina normala betydelser. Därför fordras en utredning av ordbegreppet innan det är meningsfullt att diskutera hur orden i textmaterialet bör behandlas med statistiska metoder.

2.2.1 Textord och graford

De följande avsnitten bygger till stor del på ordforskning som bedrivits under många år vid Språkdata, institutionen för svenska språket. Allmänt kan hänvisas till Nusvensk frekvensordbok 1 och 2, (Allén 1970 och 1971), inledningarna samt till Berg (1978) och Järborg (1989). En senare, mer komprimerad beskrivning återfinns i Järborg (2005).

Vid manuell analys av skriftspråk brukar man inte behöva bry sig om att de enheter som avgränsas i texten dels kan vara annorlunda än ”normala” ord, dels att även normala ord kan uppträda i olika varianter som inte har att göra med typisk böjningsvariation e.d. Vid datoriserad analys blir det däremot nödvändigt att beakta att de avgränsade textenheterna kan bestå av bland annat enstaka bokstäver (som *s.* i diverse förkortningar, t.ex. av *sida*), sifferuttryck (*2006*), vissa specialtecken (*£*), kombinationer av bokstäver och siffror (*25-åring*), för att endast nämna de vanligaste typerna. Vidare gäller, som bekant, att skiljetecken ”hängs på” det föregående eller ibland det följande normala ordet: man får alltså tekniskt olika ord *stol* beroende på om det följs av komma, punkt, kolon eller semikolon eller om det står först eller sist i en parentes. I meningsbörjan får man dessutom en variant med stor initialbokstav – versal. Annars gäller som bekant att ord inledda med versal huvudsakligen är egennamn av olika slag, vilka, enligt ovan, bör undantas från ordförrådsundersökningen. Problem uppstår dock vid egennamn som *Sten* i början av en mening; de kan av datorn inte utan vidare skiljas från det vanliga substantivet. Slutligen kan man vilja sammanföra vissa tekniskt avgränsade enheter, som uttrycket *i dag*, vilket ju varierar med den enkla enheten *idag*. Särskilt vid vissa förkortningar kan varianterna bli många: uttrycket *och så vidare* kan förkortas på minst fem sätt, av vilka vissa tekniskt är ”flerordsenheter” med mellanrum medan andra är ”ettordsenheter”. För detaljerade exempel på alla dessa typer hänvisas till kapitel 3.

Det är praktiskt att ha en beteckning för de rent tekniskt avgränsade enheter som datorn kan påträffa i texter och den naturliga termen är *textord*. Man talar även om *ortografiska ord*. Även om det stora flertalet textord givetvis utgörs av normala ord är de avvikande fallen, enligt ovan, tillräckligt många för att kunna snedvrida språkstatistiska beräkningar. Det är alltså nödvändigt att på något sätt behandla och normalisera textorden i ett material, inte minst att befria orden från påhängda skiljetecken. För denna procedur existerar dataprogram med mycket låg felprocent som brukar kallas ”tokeniserare”, eng. tokenizer. Den metod för tokenisering som används i projektet beskrivs vidare i avsnitt 3.2.3. De ”onormala” orden d.v.s. siffror, specialtecken etc., kvarstår tills vidare men kan naturligtvis nonchaleras i den fortsatta

bearbetningen, t.ex. genom att endast ord belagda i lexikon eller bestående av enbart vanliga bokstäver godtages som ingångsmaterial.

Först efter det att textmaterialet genomgått tokenisering är det meningsfullt att upprätta ordlistor över de normaliserade enheterna, vilka brukar kallas *graford*. Till frekvensen för grafordet *stol* i en sådan lista bidrager då alltså även de textord där *stol* följs av eller föregås av skiljetecken. Frekvenser över graforden i olika textmaterial är ganska enkla att erhålla och vanligen även rimligt jämförbara eftersom olika tokeniserare i praktiken ger tämligen likartade resultat. Många enklare frekvensuppgifter om ord i diverse textmaterial avser endast graforden och bör därför bedömas med viss reservation, på grund av problemen med mångtydighet, se vidare nedan. Dessutom bör man hålla i minnet att frekvenstopparna, t.ex. de hundra vanligaste orden, sällan ger något intressant utslag mellan olika material eftersom de flesta högfrekventa graforden är korta s.k. funktionsord (*och*, *i*, *att*, *det* …) som förekommer i ungefär samma proportioner i de flesta texttyper och ämnen. Allmänt gäller att de 200 vanligaste graforden täcker ca 50 % av en svensk text, naturligtvis med variation med avseende på texttyp eller genre.

Vid jämförelser av grafordsfrekvenser i olika genrer eller ämnesområden kan man på olika sätt försöka eliminera de föga utslagsgivande funktionsorden. Ett sätt är att inskränka jämförelserna till endast ord under en viss frekvenströskel (t.ex. från och med det etthundraförsta ordet). Inom projektet utfördes en delundersökning som utgick från antagandet att funktionsord vanligen är korta medan deras motsats, de s.k. innehållsorden, är långa. Med denna utgångspunkt gjordes en jämförelse mellan läroböcker i de åtta olika ämnena, avseende de mest frekventa *långa* graforden per ämne, där ”långa ord” något godtyckligt definierades som ord med minst åtta bokstäver. Det kan vara intressant att notera att denna rent mekaniska jämförelse på grafordsnivå gav mycket kraftigt utslag och att det inte vållade lekmän några svårigheter att identifiera läroböckerna/läroämnena från vilka dessa ordlistor hade hämtats. I tabell 2.1 ges exempel på dels de 10 vanligaste graforden, dels de 10 vanligaste långa graforden från en lärobok i samhällskunskap. Som synes är det ganska enkelt att bestämma ämne med hjälp av den senare kolumnen men knappast med den förra.

Tabell 2.1 Identifiering av ämne med hjälp av långa graford.

10 st. vanligaste orden i fallande frekvensordning		**10 st vanligaste ”långa” orden i fallande frekvensordning**	
Ord	*Absolut frekvens*	*Ord*	*Absolut frekvens*
och	1369	människor	108
att	1097	riksdagen	59
i	1056	företaget	57
som	971	anställda	51
är	852	regeringen	50
det	672	kommunen	42
på	588	Diskutera	40
av	572	narkotika	39
en	571	Sveriges	38
de	498	kommuner	37

Även om frekvenslistor över graford inte ger en särskilt tillförlitlig bild av ordförrådet i ett textmaterial (jfr också nedan) bör redovisningen av OrdiL-projektet givetvis innehålla grafordslistningar, och detta av två skäl. Dels är sådana listor lätta att producera rent tekniskt med mer eller mindre automatiska metoder; de förbrukar således föga resurser. Dels är det viktigt med jämförbarhet gentemot andra ordförrådsun dersökningar, vilka, som nämnts, huvudsakligen har redovisats just med grafordslistningar. Dessutom gäller att vissa av orden i fasta förbindelser endast förekommer i en och samma böjningsform, som *ögat* i uttrycket *med blotta ögat*. De fasta förbindelserna och/eller deras ingående ord måste alltså redovisas på grafordsnivå. Man skulle t.ex. i en grafordslista kunna ange att x% av beläggen på grafordet *ögat* ingår i fast förbindelse. För exempel se avsnitt 2.5.1; jfr också avsnitt 2.2.3.

Som framgått måste man räkna med att många graford är mångtydiga. Inom språkvetenskapen skiljer man mellan två huvudtyper av mångtydighet: homonymi och polysemi. Med *homonymi* brukas menas att två eller flera i grunden olika ord råkar sammanfalla i formen, som *tiger* ’stort, strimmigt kattdjur’ respektive *tiger* ’håller tyst’. Man skiljer sedan på sammanfall i uttalet, *homofoni*, och sammanfall i skrift, *homografi*; vi koncentrerar oss i fortsättningen på den senare formen. Med *polysemi* brukar, å andra sidan, menas att ett och samma ord har olika betydelser, som *färg* ’kulör’ respektive ’målningsmaterial’. Det kan naturligtvis diskuteras vad som är ”samma” eller ”olika” ord. I

exemplen kan grafordet *tiger* tydligen föras till två olika grundformer: *tiger* (substantiv) respektive *tiga* (verb) och till dessa grundformer hör inga fler böjningsformer som sammanfaller. I fallet *färg* finns däremot samma två betydelser både i grundformen och i alla böjningsformer. För enkelhets skull kan man därför basera skillnaden mellan homografi och polysemi på sådana formella egenskaper, vilket särskilt underlättar datoriserade undersökningar.

Man kan skilja mellan flera typer av homografi. Den som exemplifieras av grafordet *tiger* brukar kallas *extern* homografi, eftersom det rör sig om ett sammanfall mellan böjningsformer till olika grundord, i detta fall dessutom tillhörande olika ordklasser. Även grafordet *sköts* representerar extern homografi, men inom samma ordklass – verben *skjuta* respektive *sköta*. Ett graford som *beväpnade* kan däremot representera olika böjningsformer av samma grundord, t.ex. imperfekt som i *Han beväpnade sig med en kraftig påk*. respektive perfekt particip i plural som i *Landet har problem med beväpnade grupper*. Detta är ett exempel på s.k. *intern* homografi, vilken vanligen är mycket svårare att behandla automatiskt. Som synes är dessutom just former på *-ade* systematiskt homografa. En typ som man kanske inte tänker på kan få representeras av grafordet *and*, vilket dels kan representera ett svenskt substantiv, dels en engelsk konjunktion. Det är inte ovanligt att utländska, särskilt engelska, fraser används i en svensk text, vilket gör att man också är tvungen att beakta sådan mellanspråklig eller *interlingval* homografi. Förhoppningsvis är dock denna typ mindre vanlig i läroböcker.

Det är uppenbart att både homografi och polysemi kan vålla problem för läsförståelsen, inte minst för personer med annat modersmål än svenska. En första fråga blir då hur vanligt det är med dessa former av mångtydighet. I stort sett gäller att hälften av de löpande orden i normal svensk text är homografa, vilket är ett besvärligt problem för datorer men kanske något mindre för människor. Det är nämligen främst de korta, högfrekventa graforden som är homografa, som *för* (preposition, verbpartikel, konjunktion, substantiv, verb i presens och kanske fler alternativ), *var* (adverb, verb i imperfekt, substantiv etc.) med flera. Ofta är vissa av alternativen ganska osannolika (substantivet *skulle*, som i *höskulle*) eller är skillnaden inte så viktig för förstå-

elsen. Det spelar kanske mindre roll om *en* tolkas som räkneord eller obestämd artikel). Däremot finns ett antal mer innehållsrika graford som ofta är systematiskt homografa och kan vålla problem. En viktig typ är substantiv med plural på *-ar* som är homografa med svaga verb med presens på *-ar* (som *kammar*, substantiv eller verb). I några av de först inkodade läroböckerna förekom grafordet *hamnar* ganska frekvent; i fysikboken var det i samtliga fall fråga om verbet *hamna*, i historieboken i samtliga fall substantivet *hamn*. Många homografier kan vara en aning oväntade, som *låt* (substantiv eller verb i imperativ) och speciellt frekventa i lärobokstext av pedagogiska skäl: imperativen uppträder naturligtvis i uppgiftstext (*låt den längsta sidan i triangeln vara 15 cm*, etc.). Många fler typer av potentiellt vilseledande homografier skulle kunna anföras.

Antalet polysema ord i texter är svårare att uppskatta, delvis beroende på att det inte alltid föreligger enighet om vad som skall bedömas som olika betydelser. I många fall är det dock klart dels att det är fråga om helt olika betydelser, dels att en sammanblandning skulle kunna vålla problem. I den nämnda fysikboken hade verbet *trycka* i alla former betydelsen 'påverka genom direkt kontakt', i historieboken däremot 'mångfaldiga i tryckpress'. Denna typ av odiskutabel polysemi kan jämföras med den mer subtila typen där en konkret betydelse också har en överförd användning som i *gå några steg* vs *ett steg i utbildningen*. Det är kanske särskilt i den sistnämnda typen som andraspråksinlärarens modersmål kan interferera eftersom det ingalunda är säkert att motsvarande ord på modersmålet kan användas i liknande överförda betydelser. Jämför här diskussionen i avsnitt 2.3.2, punkt (h).

2.2.2 Upplösning av mångtydighet

I diskussionen kring homografi talades det om att graford ofta kan föras till olika s.k. grundformer. Det är dessa grundformer, t.ex. substantiv i obestämd singular, verb i infinitiv etc., som brukar förtecknas i ordböcker och ordlistor. För att en lista över orden i ett textmaterial skall bli jämförbar med orden i en ordbok fordras därför att graforden sammanförs till grupper som naturligen kan representeras av sådana

grundformer. En grupp av ord som *stol, stols, stolen, stolens, stolar, stolars, stolarna, stolarnas* brukar kallas ett *lemma*; här alltså lemmat *stol* (*-en -ar*; subst.). Ett lemma identifieras således av en grundform, en ordklassuppgift och en uppgift om böjningsmönster – paradigm. Ofta finns ett substantivlemma och ett verblemma med samma grundform: *kratta* (*-n -or*; subst.) respektive *kratta* (*-ade*; verb). Det kan också finnas flera lemman med samma ordklass men olika böjningsmönster: *rev* (*-et* plur.=; subst.) 'undervattensgrund' m. m. respektive *rev* (*-en -ar*; subst.) 'fiskelina'. I en listning av lemman i ett material kommer frekvensuppgiften för *stol* (*-en -ar*; subst.) således att sammanfatta frekvenserna för alla de åtta formerna ovan. Uppenbarligen ger frekvensuppgifter för lemman en bättre bild av ordförrådet än frekvenser för graford. Numera finns också möjligheter att med automatiska metoder föra graford i en text till rätt lemma med ganska god säkerhet.

Sammanförande av olika böjningsformer till lemman brukar, med en naturlig term, kallas *lemmatisering.* För denna procedur fordras ett lexikon som för "samtliga" svenska lemman kan ange samtliga möjliga böjningsformer, antingen direkt som en slags lista eller indirekt genom ett system av böjningsregler för varje paradigm. För korrekt resultat fordras att textmaterialet, med hjälp av ett dylikt lexikon, först uppmärks med alla möjliga alternativ, s.k. *homografkomponenter*, för varje graford. Därefter måste texten genomgå *homografseparering*, innebärande att det riktiga eller sannolikaste alternativet för varje homograft graford utväljs. Homografseparering i större skala utförs numera maskinellt med hjälp av programsystem för s.k. ordklasstaggning, vilka förser graforden med uppgift om ordklass och böjningskategori med ledning av motsvarande uppgifter för närstående ord. Om man t.ex. betraktar det homografa ordet *kammar* i satserna *Hennes kammar ligger i lådan* och *Hon kammar sig ofta* kan följande analys utnyttjas:

Hennes	*kammar*	*ligger*	*i*	*lådan*
poss. pron.	?	verb	prep.	subst.
sing	?	pres.	-	best. sing.

Hon	*kammar*	*sig*	*ofta*
pron. 3:e p	?	refl. pron.	adv.
grundform	?	3:e p.	-

Här kan man se att *kammar* måste vara substantiv i ställningen mellan ett possessivt pronomen och ett finit verb, liksom det måste vara verb mellan ett personligt pronomen i grundform och ett reflexivt pronomen. En ordklasstaggare innehåller en stor mängd dylika regler, ofta automatiskt upprättade från ett manuellt uppmärkt utgångsmaterial. Det lexikon och den ordklasstaggare som används i OrdiL-projektet presenteras kortfattat i kapitel 3.

Även om homografseparering och lemmatisering i det övervägande antalet fall ger korrekt resultat kan en viss felprocent, ca 3 %, aldrig undvikas. Innan ordförrådet presenteras i olika lemmalistor är det därför nödvändigt att försöka korrigera åtminstone de mest elakartade felmärkningarna manuellt. Eftersom huvuddelen av informationen i en text uppbärs av s.k. innehållsord (substantiv, verb, adjektiv) bör korrigeringen i första hand inriktas på klart vilseledande felmärkningar som rör dessa ordklasser, t.ex. ord som *låg*, verb eller adjektiv; jfr även *hamnar* m.fl. För sådan selektiv korrigering fordras viss erfarenhet och språklig fantasi, så att inte de manuella insatserna förslösas på onödiga kontroller, t.ex. av osannolika homografier, som *skulle* ovan. Av arbetstekniska skäl kombineras lämpligen korrigering av lemmatiseringen med den semantiska uppmärkningen av polysema ord enligt det följande.

Eftersom många lemman, särskilt de frekventa, har flera betydelser ger inte heller frekvenslistor över ett lemmatiserat ordförråd en helt rättvisande bild av en text. Den bästa lösningen vore därför att alla ord i läroboksmaterialet kunde uppmärkas semantiskt. Eftersom denna lösning är orealistisk kan man tänka sig en kompromiss, innebärande att alla lemman med flera betydelser (lexem) får en särskild markering i listorna som ett slags varningssignal: Observera flera betydelser! En sådan markering kan i princip automatiskt inhämtas från den lexikala databasen GLDB, se nedan. Lemmat *trycka* (*-te*; verb) skulle alltså kunna markeras med på lämpligt sätt, t.ex. med fetstil, asterisk eller en siffra för antalet möjliga betydelser; för exempel se 2.5.3.

Den mest önskvärda enheten för en ordförrådsundersökning är tydligen vad som hittills kallats en klart urskiljbar ”betydelse” av

ett lemma. För sådana enheter behövs en term, nämligen *lexem*. För det polysema lemmat *trycka* (*-te*; verb) kan man etablera fyra lexem, nämligen:

trycka /1 'påverka genom direkt kontakt i riktning från kraftkällan' *gasen trycker mot kärlets väggar* etc.
trycka /2 'bekymra' *vad är det som trycker dig?* etc.
trycka /3 'mångfaldiga i tryckpress' *förlaget lät trycka boken i 5000 exemplar* etc.
trycka /4 'hålla sig dold och orörlig' *haren låg och tryckte* etc.

Av dessa förekommer, som sagts, enbart *trycka/1* i den undersökta fysikboken medan *trycka/3* förekommer i historieboken. Det är naturligtvis svårare att skilja mellan olika lexem än mellan olika lemman eftersom lexemen måste etableras genom en ibland ganska komplicerad betydelseanalys. I praktiken behöver man därför ett lexikaliskt underlag för semantisk uppmärkning. Lexemen för *trycka* har hämtats från GLDB (Göteborgs lexikala databas) och de semantiska beskrivningarna är förkortade och förenklade versioner av GDLB:s definitioner. För en allmän presentation av GLDB se t.ex. Järborg (2003a). Databasen är mest känd som det tekniska underlaget till en serie ordböcker, först Svensk ordbok (1986), senast Nationalencyklopediens ordbok (1995-96); en ny version väntas om några år. För OrdiL har GLDB två fördelar: dels att den är tekniskt lättillgänglig för uppmärkning, dels att det finns erfarenheter av att uppmärka textmaterial med dess lexem, nämligen från projektet Lexikalisk betydelse och användningsbetydelse, Järborg (2003b). Vidare är den semantiska analysen i GLDB förmodligen mer djupgående än i alternativa lexikala verk. Därför är det självklart för OrdiL-projektet att för sin partiella semantiska uppmärkning använda GLDB som underlag.

Semantisk uppmärkning är nödvändigtvis manuell och mycket tidsödande. Försök har gjorts med automatisk semantisk uppmärkning men även i de mest lyckade experimenten, baserade på kraftfullt manuellt förarbete, når man inte mer än 67 % rätt. Detta resultat var f.ö. segrarnoteringen i "tävlingen" SENSEVAL 2 i automatisk semantisk uppmärkning, vilken använde material från projektet Lexikalisk be-

tydelse och användningsbetydelse. Se vidare Kokkinakis, Järborg & Cederholm (2001). Det kan alltså aldrig bli fråga om att märka hela textmaterialet eller ens alla ord i huvudordklasserna. Strategin får i stället vara (i) att märka upp polysema lemman i frekvensordning, (ii) att bland huvudordklasserna prioritera i första hand substantiven, i andra hand verben eftersom substantiven normalt har mer specifik betydelse, (iii) att prioritera klart vilseledande polysemier, d.v.s. semantiskt starkt skiljaktiga lexem, före mer subtila, (iv) att lågprioritera polysemier som involverar klart sällsynta eller marginella lexem, som *pistol/2* 'ett slags guldmynt'. Dessutom måste man vid den semantiska uppmärkningen tyvärr nonchalera överförda eller utvidgade s.k. *underbetydelser* till lexemen. T.ex. har *trycka/1* en underbetydelse 'påverka mentalt till att godtaga', som i *han tryckte särskilt på att de måste inkludera jämställdhetsaspekten*. Denna underbetydelse får vid uppmärkningen anses vara härledbar från 'påverka genom direkt kontakt', den s.k. *kärnbetydelsen*, men att konsekvent markera även underbetydelser skulle vara alltför tidsödande.

Som nämnts ovan är det naturligt att samordna den semantiska uppmärkningen med den manuella kontrollen av lemmatiseringen. För båda momenten fordras viss lexikalisk kunskap och erfarenhet, så att endast fall som både är frekventa och potentiellt klart vilseledande kommer att behandlas. I praktiken blir det fråga om att utgå från frekvensordnade grafordslistor, lämpligen på papper, och märka upp alla intressanta fall, vilka sedan kontrolleras och rättas, respektive märks upp, på datorskärmen. Med denna strategi bör de fall som inte medhinns alltså vara lågfrekventa och/eller föga besvärande för förståelsen. Det program som används för manuell kontroll och uppmärkning, liksom det tekniska formatet hos den underliggande databasen GLDB, beskrivs och illustreras i kapitel 3.

Medan det torde vara möjligt att presentera mer eller mindre fullständiga lemmalistor enligt ovan, åtminstone för huvdordklasserna, kan den semantiska uppmärkningen endast resultera i listor över lemman med särskilt intressanta lexem; d.v.s. lemman som är mångtydiga, potentiellt vilseledande och vars lexem är tillräckligt frekventa. Varje lexem i listningarna måste, förutom uppgifter om frekvens, spridning

o.d., också förses med en identitetsuppgift och/eller en enkel definition. Det kan kanske vara lämpligt att lista de definierade lexemen separat eftersom definitionerna ofta fordrar ett visst utrymme. I de vanliga listorna kunde man således skriva *tryck/3*, med uppgift om frekvens och spridning, och i en separat lista ange definitionen, i praktiken kanske i form av en förenklad variant av definitionen i GLDB. I frekvenslistorna kunde man också förse vissa lexem med en markering att de innefattar vanligen abstrakta underbetydelser, analogt med den motsvarande varningssignalen för lemman med flera lexem. Lexemet *tryck/1* skulle, enligt ovan, fordra en sådan markering. För exempel hänvisas till 2.5.3 och 2.5.4.

2.2.3 Flerordsenheter och morfemfamiljer

I stora textkorpusar brukar man kunna finna relativt många flerordsuttryck som återkommer. En del av dessa är högfrekventa men ganska ointressanta, som typen *Det är* (i meningsbörjan), *och att*, *har inte* och många liknande. Å andra sidan kan man påträffa vissa s.k. idiom, som *kasta/kastar/kastade yxan i sjön*, vilka är språkligt intressanta men lågfrekventa. Eftersom idiom kan vara okända för ovana läsare och har en helhetsbetydelse som definitionsmässigt inte kan härledas från de ingående ordens betydelser bör de rimligtvis i stort sett saknas i lärobokstext, något som bekräftas av iakttagelser i OrdiLs lärobokskorpus. I den mån idiom dock påträffas är de naturligtvis av särskilt intresse och hela kategorin kunde vara värd en egen undersökning. Mellan dessa extremfall kan man emellertid finna många ganska frekventa flerordsuttryck som visserligen inte är helt oförutsägbara men dock skulle kunna bereda tolkningssvårigheter. Undersökningen av förrådet av enstaka ord i lärobokskorpusen blir sålunda inte representativ om den inte kompletteras med en undersökning av vissa typer av flerordsenheter.

Vissa flerordsuttryck igenkänns redan av den tokeniserare som används i projektet och behandlas därefter som om de vore ettordsenheter. Detta gäller främst den typ som kan kallas ”utbyggda preposi-

tioner", som *på grund av*. I en förundersökning, utförd på en av läroböckerna i fysik, framkom några andra typer som kan anses potentiellt svårtolkade och alltså bör redovisas särskilt. Till dessa typer hör (i) flerordsnamn som *norra halvklotet*, (ii) namnliknande uttryck, som *absoluta nollpunkten*, *den heliocentriska världsbilden* (Obs. bestämd form!), (iii) flerordstermer, som *bunden rotation*, *elektriska ledare*, *gröna växter* (ej helt förutsägbara betydelser!), (iv) allmänna fraser som *med blotta ögat* 'utan optiska hjälpmedel', (v) flerordsverb som *lägga märke (till)*, *ta reda (på)*, (vi) ytterligare flerordsprepositioner, som *i form av*. Det kan naturligtvis diskuteras huruvida alla fall av dessa typer verkligen är svårtolkade och i sista hand får därför varje flerordsuttryck bedömas för sig. Det finns dock anledning att i tveksamma fall hellre registrera uttrycket eftersom det är möjligt att också redovisa de ingående orden separat, se nedan.

Urvalet av flerordsenheter måste naturligtvis göras manuellt. Utgångspunkten bör vara att undersöka flerordsuttryck som uppträder "påfallande ofta" i lärobokskorpusen. För operationalisering av "påfallande ofta" existerar ett antal statistiska metoder, vilka i princip mäter och värderar förhållandet mellan flerordsuttryckets frekvens och de ingående ordens frekvenser vart för sig. Om orden förekommer tillsammans betydligt oftare än vad en slumpmässig fördelning skulle ge får man alltså ett mått på "statistisk samhörighet", vilket givetvis kan vara en indikation på språklig samhörighet. För OrdiL har valts det s.k. log-likelihoodmåttet, se vidare kapitel 3. Med hjälp av detta mått produceras en lista över flerordsuttryck, två- eller treordsuttryck, ordnade efter statistisk samhörighet och med frekvens angiven, ur vilken de intressanta flerordsenheterna sedan utväljs, enligt de principer som antytts ovan. Genom denna arbetsgång säkerställs att inga frekventa eller statistiskt starkt samhöriga flerordsuttryck utelämnas.

De valda flerordsenheterna bör redovisas med frekvens, spridning etc. i särskilda listor. Det kan diskuteras huruvida redovisningen bör ske i anslutning till grafordslistorna eller till lemmalistorna. I allmänhet är det inte rimligt att lemmatisera orden i flerordsenheterna, som påpekats ovan; i frasen *med blotta ögat* är det således grafordet *ögat* och inte lemmat *öga* (*-t, ögon*; subst.) som är aktuellt. Däremot finns det

anledning att på lämpligt sätt lemmatisera just verbet i flerordsverb, så att former som *lägga/lägger/lade/lagt märke till* kan sammanföras och erhålla en gemensam frekvenssiffra. Givet att varje flerordsenhet är markerad i texten är det också möjligt och lämpligt att ange i de vanliga grafordslistorna frekvensen för ett ord i flerordsenheter, att jämföras med dess totalfrekvens. För t.ex. grafordet *märke* kunde då anges: Frekvens totalt i OrdiL 47, i flerordsenheter 24. Se avsnitt 2.5.1.

Lemmatisering innebär, som framgått, att olika böjningsformer sammanförs. En näraliggande tanke är att sammanföra även avledningar till samma grundord eller grundmorfem till s.k. ordfamiljer, något som gjorts i den kända Academic Wordlist, se kapitel 1. Motiveringen för detta är naturligtvis att kunskap om ett grundord, t.ex. *oxid* rimligen förstärks och kompletteras av ord som *oxidera*, *oxidation* etc. I ett sammansättningsrikt språk som svenska skulle man även kunna inkludera sammansättningar som *koldioxid*, *oxidationsprocess* etc. Problemet är att ordfamiljer måste upprättas huvudsakligen på manuell väg; om sammansättningar skall inkluderas är det inte heller klart var man skall göra halt i urvalet. Vidare kan det diskuteras hur mycket betydelsen skall få förändras inom en ordfamilj: *oxidation* används numera i kemiska sammanhang i betydelsen 'bortförande av elektroner från visst ämne' i stället för den ursprungliga 'bildande av oxid av visst ämne'. Frågan om ordfamiljer är ännu inte slutdiskuterad inom OrdiL men mycket talar för att åtminstone några välgrundade ordfamiljer bör kunna redovisas i särskilda listor. Även utan en sådan redovisning kan man dock notera att avledningar till ett givet grundord normalt uppträder i omedelbar anslutning till detta i en alfabetiskt ordnad lista. För användaren av listan är det följaktligen möjligt att, med lite besvär, själv räkna ut frekvensuppgifter för många ordfamiljer. Vissa tekniska metoder för sammanförande av böjningsformer existerar redan, se kapitel 3, och dessa metoder kunde kanske utvecklas till att hantera även avledningar.

2.3 Lexikal kompetens i praktiken

Den allmänna kognitiva och sociala betydelsen av ett stort ordförråd och goda ordkunskaper hos skolelever liksom hos medborgare i

allmänhet har alltid varit känd men forskning kring ordförrådets storlek och kvalitet har, av naturliga skäl, varit betydligt mer utbredd inom pedagogisk och tillämpad språkvetenskap än inom den rent teoretiska lingvistiken. För sådan forskning är det givetvis centralt att kunna bedöma kvantitet och kvalitet i enskilda språkbrukares ordkunskaper. För sådana bedömningar fordras dels en modell för lexikal kompetens, dels ett lexikalt underlag som är i någon mening representativt, dels en uppsättning mätmetoder, vilka täcker alla modellens dimensioner av lexikal kompetens. Eftersom ett av målen för OrdiL-projektet är att skapa ett representativt lexikalt underlag för bl.a. ordkunskapstester och, i förlängningen, även utprova dylika; se kapitel 4 och 5, finns anledning att diskutera hur de kvantitativa och kvalitativa aspekterna av lexikal kompetens kan yttra sig i praktiken.

2.3.1 Tidigare lexikala modeller

För en bredare översikt av ordförrådsforskning hänvisas till kapitel 1. Här skall endast några viktiga modeller för lexikal kompetens kort diskuteras, delvis med utgångspunkt i de teoretiska resonemangen kring det mentala lexikonet och ordbegreppet i avsnitten 2.1 och 2.2. Av de modeller som nämnts i det föregående tycks Stroud (1979), Nation (2001) och Qian (2002) vara de som närmast anknyter till OrdiL-projektet. Strouds modell har tillkommit inom ramen för svensk andraspråksforskning och är därigenom givetvis av speciellt intresse i sammanhanget. Nations modell har varit den internationellt kanske mest inflytelserika, bl.a. genom att den ligger till grund för välkända ordförrådstest som *VLT*, se kapitel 1. Dessa modeller, vilka har mycket gemensamt (Strouds är dock mer detaljerad) inriktas huvudsakligen på lexikonet som ett statiskt system. I motsats härtill kan Qians modell sägas i högre grad inkludera dynamiska aspekter, lexikonet i funktion; sådana aspekter är naturligtvis centrala för studiet av läsförståelse, ordinlärning m.m.

Ett generellt problem med många dylika modeller, inklusive de nämnda, är en svävning i användningen av det vaga begreppet ”ord”. Ibland betraktas detta enbart som en uttrycksenhet, t.ex. en

följd av bokstäver, ibland som en union av uttryck och innehåll, en bokstavskombination *plus* en s.k. betydelse, eventuellt flera. Även på uttryckssidan kan det finnas svävningar mellan textord, graford och lemman, se avsnitt 2.2, särskilt i modeller utvecklade för engelska, på grund av detta språks enklare böjningsmorfologi. Det borde vara självklart i ett perceptions- och avkodningsperspektiv att det inte kan vara fråga om verklig ordigenkännelse utan någon form av tolkning varför den naturliga modellen för den lexikala enheten måste vara ett *par* bestående av ett uttryck och ett innehåll, d.v.s. i princip ett ”tecken” enligt de Saussure. Omvänt har alltså varje lexikal enhet dels en uttryckssida, dels en innehållssida, där den senare givetvis är mycket mer komplex. Som framgått av 2.2 kan den grundläggande enheten kallas *lexem* och uttrycksenheten *lemma*, vilken senare sammanfattar en grupp böjningsformer.

I de nämnda modellerna för lexikal kompetens avspeglas osäkerheten kring ordbegreppet i den skenbart lika värderingen av de antagna lexikala dimensionerna. Sålunda antager Nation åtta dimensioner, av vilka endast två kan sägas vara direkt innehållsliga. Dessa dimensioner tycks betraktas som lika tungt vägande som ordets stavning och uttal och även som en rimligtvis sekundär dimension som ordets ”frekvens”. Ingen antydan ges hos Nation eller Stroud om beroendet mellan ett ords innehåll och dess grammatiska och kollokationella egenskaper. Därvid anknyter de dessvärre till en gammal lingvistisk tradition att betrakta de systematiska komponenter som brukar användas i pedagogiska sammanhang, fonologi, morfologi, syntax etc., som om de hade en oberoende psykolingvistisk eller kognitiv realitet, något som återstår att bevisa. Några sådana beroenden antyds däremot i Qians modell i den mycket allmänna beskrivningen av den fjärde dynamiska lexikala dimensionen, innefattande olika lexikala processer. Tyvärr är dock Qians beskrivning tämligen abstrakt och därför svåranvänd i praktiska tillämpningar.

2.3.2 En konkret modell för lexikal kompetens

Detta avsnitt kan ses som en praktisk applikation av de tidigare teoretiska resonemangen. Tonvikten har lagts på sådan aspekter av

lexikal kunskap som är kritiska dels för språkreception särskilt av text, dels för egen produktion. En bärande tanke är att kunskap om ordens innehållssida är primär och att mycket kunskap som traditionellt kallas ”grammatisk” kan anknytas mer eller mindre direkt till ordens betydelser. Följaktligen beskrivs inga egentliga oberoende ”dimensioner”; tonvikten läggs i stället på rent funktionella lexikala kunskaper.

(a) Kunskap om själva uttrycksformen, både ljudbild och särskilt skriftbild, är en förutsättning för all reception och alltså för de följande punkterna. För receptionen kan långa ord vålla svårigheter; se vidare punkt (d) och (e). Bristande förståelse för sambandet mellan ljudbild och skriftbild skulle eventuellt kunna försvåra en morfologisk analys av okända ord. Uttrycksmässig likhet eller överensstämmelse, homografi, homofoni, med böjningsformer från andra lemman, jfr fallet *hamnar* i 2.2, kan inte så sällan leda till störningar i receptionen. För undvikande av sådana störningar fordras kunskap om böjningsmönster hos uttrycksmässigt liknande ord, jfr även (b), och gärna ett slags lexikal fantasi, besläktad med förmåga till s.k. ordvitsande. Osäkerhet om stavning eller uttal kan utgöra hinder för produktiv användning och kanske leda till överanvändning av en inadekvat men lättstavad närsynonym.

(b) Särskilt för receptionen är det viktigt att kunskap om uttrycksformen omfattar hela lemmat, d.v.s. samtliga böjningsformer. Oregelbundna paradigm, *bära-bar-burit* etc., är naturligtvis det stora problemet. För produktionen kan noteras att det brukar vara fullt möjligt att göra sig förstådd med regelbunden böjning, *bära-bärde*, även om acceptansen hos lyssnarna för dylika ”språkfel” kan variera med ålder och bakgrund.

(c) Medvetenhet om att ett lemma är polysemt, d.v.s. kan representera flera lexem, är absolut nödvändig för tolkningen, särskilt i de fall där lexemen är semantiskt obesläktade, som *skjuta/1* ’sända iväg projektil’ respektive *skjuta/5* ’förflytta genom direkt tryck’. Detta fordrar ganska bred, men ej nödvändigtvis djup, lexikal kunskap; med andra ord att det kan räcka att känna till att det finns flera lexem och ha viss uppfattning om deras ungefärliga betydelser.

(d) För receptionen måste vanligare avledningsändelser, *-lig*; *-het*; *-bar*; *-ning/-ing*; *-(is)era*; *-(t)ion*, och deras betydelser vara kända, t.ex. att *-het* lagd till ett adjektiv betyder ungefär 'egenskap bestående i att vara ADJEKTIV'. Motsvarande gäller för vissa vanliga förled, t.ex. *o-* vid adjektiv. För produktionen är det viktigt att korrekt kunna uttrycka sådana abstraktioner över konkreta begrepp så att man kan bilda t.ex. *undersökning* 'handling bestående i att undersöka något' från *undersöka*.

(e) För receptionen måste tillfälliga sammansättningar kunna analyseras i korrekta delar, vanligen två, vilket på uttryckssidan fordrar kunskap om möjliga stavelsegränser och dyligt och ibland om speciella sammansättningsformer, *kyrko-*, *rosen-*. Väsentligare är igenkänning av leden som möjliga lexem och rekonstruktion av den troliga semantiska relationen mellan dem, *silversked* 'sked som består av silver'; *soppsked* 'sked som används till soppa'; dessa moment fordrar framför allt kunskap på innehållssidan enligt (g) nedan.

(f) För receptionen fordras en viss beredskap att inte nöja sig med identifiering av enstaka ord utan vara medveten om att de kan ingå, ofta med helt annan betydelse, i flerordslexem. Man kan exemplifiera med tvåordsverbet *lägga märke (till)*, där inget av de vanliga lexemen under *lägga* repektive *märke* är aktuella. Naturligtvis fordras också kunskap om själva tvåordsenheten i sig. Flerordsenheter kan ha mycket olika status, från mera kulturellt betingade (*röda stugor*), över historiska uttryck (*hank och stör*), partikelverb (*tycka om*) och andra flerordsverb, flerordsprepositioner (*på grund av*) och fram till s.k. idiom (*sopa något under mattan*). Den sistnämnda typen är givetvis särskilt besvärlig för inläraren.

(g) Den centrala innehållsliga kunskapen om ett lexem kan beskrivas genom en strukturerad analys i s.k. betydelsefaktorer, se Järborg (2003a); t.ex. att *måla* i ett av lexemen innebär att någon person anbringar ett trögflytande material, färg, som har viss kulör, på ett föremål, vanligen med hjälp av pensel, roller e.d., i syfte att pryda och/eller skydda föremålet. Som framgår omfattar en så pass fullständig analys som denna ganska många och detaljerade betydelsefaktorer, vilket implicerar att man kan vänta sig att det

kan existera olika stadier av partiell innehållslig kunskap med färre eller mindre detaljerade betydelsefaktorer. För full förståelse av ett lexem fordras naturligtvis detaljerad innehållslig kunskap men sådan kunskap är också viktig för en differentierad och mer exakt egen språkproduktion, jfr Viberg (2004).

(h) Givet en grundläggande kunskap om ett lexems väsentliga innehåll, den s.k. kärnbetydelsen, måste man även, särskilt för fullständig textförståelse, ha kunskap om semantiken bakom utvidgade eller överförda användningar, s.k. underbetydelser. Ofta finns en konkret kärnbetydelse, som i *hon sköt över ID-kortet till kassörskan*, till vilken hör en abstrakt underbetydelse, som i *han sköt över ansvaret på sin revisor.* Olika språk kan variera kraftigt i fråga om vilka centrala lexem som kan ha abstrakta underbetydelser. Om man t.ex. betraktar centrala (själv)förflyttningsverb i svenska respektive engelska finner man att underbetydelsen 'vara i funktion om mekanism e.d.' finns vid verbet *gå* på svenska, *klockan går*, men vid verbet *to run* på engelska, *the watch is running.*

(j) Semantiska relationer i lexikonet brukar, efter de Saussure (1916), indelas i paradigmatiska och syntagmatiska. Bland de paradigmatiska har man traditionellt uppmärksammat sådant som över- under- och sidoordning, som att *häst* är underordnat *hovdjur*, överordnat *ardenner* och *gotlandsruss* och sidoordnat med *åsna* och *zebra*. Här tillhör *häst* den s.k. basnivån och bör därför vara mer förtroget än de högre eller lägre nivåerna. Andra paradigmatiska relationer som brukar nämnas är synonymi, äkta sådan är dock sällsynt, och antonymi, motsatsord; aktuellt mest för adjektiv. För receptionen, särskilt för partiell förståelse, är kunskap om åtminstone de vanligare över-/underordningsrelationerna viktig. T.ex. bör man åtminstone känna till att en *ardenner* är ett slags *häst*. I själva verket kan just sådan kunskap utgöra den minimala formen av innehållslig kunskap om ett lexem. Även de motsatsord som utgör ändpunkter på en skala kan vara viktiga, *ljus* vs *mörk* etc.

(k) Vissa typer av syntagmatiska relationer är givna redan genom detaljkunskap om lexemets innehåll, som i exemplet från (g) där *måla/1*

uppenbarligen är relaterat till *färg* (*färg/2*) och *kulör* samt *färg/1*, liksom till *pensel*, *roller* m.m. I själva verket kan dylika relationer utvecklas till ett slags lexikalisk-semantiskt nätverk med oklara gränser, här ett ”ytbehandlingsnätverk”, där även alternativa processer som *slipa*, *polera*, alternativa redskap som *sandpapper*, typiska föremål som *plank*, *husvägg*, *köksstol* kan ingå. Sammanfattningsvis kan man säga att varje någorlunda innehållstungt lexem även bestämmer en förväntad *semantisk kontext* som man kan känna till i större eller mindre utsträckning. Denna förväntade kontext är viktig både för förståelse, t.ex. av flertydiga ord, och för egen differentierad produktion. Observera att så snart målning är introducerad kan man nämna *färgen*, *penseln* etc. i bestämd form, trots att själva orden inte har använts tidigare i texten. Om man känner till relationsnätverket kring *måla/1* behöver man alltså inte bli förvånad över att *penseln* plötsligt dyker upp.

(l) Till den semantiska kontexten i vidare mening hör även de svaga betydelsefaktorer som brukar kallas konnotationer. Ett lexem som *midsommar* är för de flesta svenskar förknippat inta bara med den tid på året då solen står som högst utan också med en fest med speciell mat och dryck, kanske speciella danser och sånger och inte minst med romantik och erotik. Just dessa konnotationer är både allmänna och starkt kulturberoende och därför ofta svårförståeliga för andraspråksinlärare. För att man skall kunna tillägna sig bl.a. skönlitterär text är kunskap om sådana konnotationer givetvis viktig.

(m) I denna modell betraktas kunskap om ett lexems ordklass och övriga syntaktiska egenskaper som sekundär till kunskapen om dess innehåll. Vid läsning identifierar man inte ett grafords ordklass separat utan genom att känna igen uttrycksdelen av ett lexem, eventuellt flera. Däremot fordras kunskap om de syntaktiska ledtrådar med vars hjälp man kan identifiera lexem i deras kombinationer med andra lexem, eller omvänt, koda en begreppslig kombination så att den blir förståelig för lyssnaren eller läsaren. Om man jämför *han sade en sak till henne* och *han berättade en sak för henne* ser man att i det förra fallet kan ”adressaten” bl.a. uttryckas med en prepositionsfras med *till* och i det senare fallet bl.a. med en prepositionsfras med *för*. Vid *säga* kan också ”ämnet”

uttryckas med direkt citat, som i *han sade "jag vet inte"*, vilket inte gäller för *berätta*. Denna typ av kunskap beskriver alltså samspelet mellan de lexikala enheternas semantik och deras syntaktiska uppträdande. Sådan kunskap är avgörande både för identifiering av rätt lexem i vissa fall, jfr *intresse för något* och *intressen i något*, och framför allt för förståelsen av de semantiska sambanden i text.

2.4 Vokabulärer i text

I avsnitt 2.1 diskuterades lexikal kunskap från ett teoretiskt perspektiv. I detta avsnitt blir det fråga om att konkretisera de teoretiska resonemangen genom att antaga att det finns vissa typer av delvokabulärer i lärobokskorpusen, vilka kan förmodas vara kända i olika grad av läsarna/skoleleverna. Nästa steg blir sedan att med hjälp av mätbara egenskaper, som enheternas frekvens och spridning, operationellt definiera dessa delvokabulärer, så att de mer eller mindre automatiskt kan identifieras i textmaterialen.

2.4.1 Vokabulärkategorier

Vid undersökning av vokabulärer från en inlärarsynpunkt kan det vara ändamålsenligt att uppdela ett textmaterials ordförråd i delvokabulärer, trots att det teoretiskt knappast finns skarpa gränser mellan delarna. För OrdiL blir det i första hand viktigt att skilja mellan *ämnesneutrala ord* ((A) och (B)) och *ämnesrelaterade ord* ((C) och (D)), i andra hand att upprätta relevanta distinktioner inom dessa överordnade kategorier. Vi erhåller då följande:

(A) *Allmänspråkliga, frekventa ord.* Härmed menas alltså ord som kan förekomma både i tal och skrift samt i både vardagliga och mer formella sammanhang. Exempel: *ha*, *vara*, *komma*, *människa*, *exempel*, *på*, *många, stor.*

(B) *Allmänna ofta abstrakta skriftspråkliga ord.* Härmed menas ord som uppträder (med lägre frekvens) i ordinär sakprosa, t.ex. i typisk

tidningstext och i de flesta läroböcker och liknande utredande texter. Betydelserna är relativt generella och normalt användbara för beskrivningar av abstrakta sammanhang, argumentation o.d. Kategorin motsvarar delvis den etablerade engelska kategorin ”academic words”. Exempel ur materialet: *utbredning*, *resurser*, *bilda*, *framträdande*, *avta*, *påverka*, *motsvara*, *föremål*.

(C) *Allmänspråkliga ämnestypiska ord*. Härmed menas ord som relativt frekvent uppträder i läroböcker o.d. i ett visst ämne eller en grupp av ämnen. Kategorin kan uppenbarligen fingraderas: man kan tala om ämnestypiska ord för naturvetenskaper i allmänhet NO-typiska ord, ämnestypiska ord för fysik och kemi, ämnestypiska ord för fysik, etc. Sådana ord behöver inte vara facktermer utan kan vara tämligen ”vanliga” ord som dock råkar vara minde vanliga i allmänspråket, trots att deras betydelser mycket väl kan vara relevanta i allmänna sammanhang. Just dessa ”vanliga” ämnestypiska ord, d.v.s. ämnestypiska ord, med uteslutande av kategorin (D) nedan, har i Golden & Hvenekilde (1983) visats kunna vara oväntat besvärliga för andraspråksinlärare. Se även diskussionen i kapitel 4 om vokabulärtest. Exempel ur materialet, ett från vart ämne: *arbetare*, *klimat*, *sträckan*, *blandning*, *strålning*, *muskel*, *företag*, *kyrka(n)*.

(D) *Fackord och facktermer, ofta unika för ett visst ämne*. Härmed ord som i princip endast förekommer i läroböcker eller motsvarande och som har betydelser vilka knappast är relevanta utanför ett visst ämne eller en viss grupp av ämnen. Som framgår av beteckningen skulle man principiellt kunna dela upp denna kategori i två. Eftersom det, i de flesta vetenskaper, sällan råder någon enighet om vad som skall betraktas som verkliga facktermer förefaller det lämpligast att här sammanhålla kategorin, särskilt som det är nödvändigt att upprätta den med formella och/eller statistiska metoder, se avsnitt 2.4.3. Exempel ur materialet, ett från vart ämne: *produktionsfaktorer*, *reformationen*, *barrskogsbältet*, *tidig-modern*, *decimalform*, *kopplingsschema*, *elektrolyt*, *kromosom*.

Den underliggande hypotesen är naturligtvis att ord i kategorierna (B) och (C) brukar kunna återfinnas i de yttre skikten av det mentala lexikonet hos en andraspråksinlärare, enligt figur 2.1, och alltså kan

vara bekanta men föga förtrogna. Ord ur kategori (D), däremot, antages normalt vara obekanta i ett initialskede och alltså inte ingå i det mentala lexikonet.

2.4.2 Material för vokabulärundersökningen

De olika vokabulärkategorierna ovan beskrevs i termer av ”typiskt” och ”otypiskt”, vilket givetvis måste tolkas i förhållande till språkliga material. OrdiL-projektet arbetar med tre större språkliga undersökningsmaterial eller *korpusar*: dels, och huvudsakligen den speciella läromedelsskorpusen, OrdiL-korpusen, nedan kallad enbart OrdiL, dels två allmänspråkliga kontrollkorpusar, nedan kallade PRESS respektive UNGDOM. Korpusarnas tekniska uppbyggnad och bearbetning m.m. beskrivs i detalj i avsnitt 3.2. Här skall närmast deras språkliga karaktär och användning diskuteras.

Läromedelskorpusen har hämtats från 16 allmänt använda läroböcker för grundskolans årskurs åtta. Kunskapsstoffet för denna näst sista årskurs kan antagas vara bekant för de allra flesta elever som går ut den obligatoriska skolan. De 16 böckerna fördelar sig på 8 ämnen med två böcker i vardera ämnet. Läroböcker i språk uteslöts på grund av att de, förutom vissa grammatiska termer, huvudsakligen bör innehålla genreneutralt svenskt språk – innehållet måste ju vara allmänt och lättförståeligt. Minst hälften av läroböcker i främmande språk består dessutom av text på just detta språk. Ämnena utgörs således av matematik, fysik, kemi, biologi, d.v.s. NO-ämnen inklusive matematik; geografi, historia, samhällskunskap, religion, d.v.s. SO-ämnen. Korpusen omfattar knappt 1 miljon ord. Vid inläsningen har inte bara själva orden utan också textens typografi och layout registrerats på lämpligt sätt, se avsnitt 3.3, vilket är viktigt för den fortsatta bearbetningen. Korpusen måste rimligen anses vara representativ för ett läromedelsspråk som alla skolelever förr eller senare kommer i kontakt med.

De allmänspråkliga korpusarna erbjuder egentligen ett större problem eftersom det ingalunda är klart vad som skall förstås med ”allmänspråk”. Till en början måste man göra en skillnad mellan

talspråk och skriftspråk. Inom talspråket kan man dessutom skilja mellan å ena sidan ”vardagligt talspråk” som i sin tur sönderfaller i språk i hemmet, på skolgården, på arbetsplatsen etc., å andra sidan formellt talspråk som i TV och radio, särskilt i nyhetsprogram o.d. På liknande sätt kan man skilja mellan olika typer av skriftspråk, från lagtexter och tekniska bruksanvisningar till annonser och löpsedlar. Den typ av skriftspråk som kanske är mest allsidig och representativ brukar antagas vara tidningsspråk, vilket också är språket i referenskorpusen PRESS2000. Det är naturligt att i första hand utnyttja tidningsmaterial, dels på grund av själva storleken, dels på grund av att tidningsspråk får antagas vara någorlunda välbekant för de flesta läskunniga, dels också för att tidningsspråk står nära det formella talspråket. Det är dock välkänt att tidningsspråk är mer abstrakt än man kunde vänta och att i synnerhet konkreta, vardagliga substantiv (s.k. tandborstord) är anmärkningsvärt lågfrekventa, vilket man får korrigera för vid olika jämförelser. Den tilltänkta läsargruppen för läroböckerna, nämligen skolelever i årskurs åtta (15-åringar), kan nog antagas i viss utsträckning vara bekant med tidningsspråk, åtminstone i uppläst form (som i TV-nyheter etc.). Däremot är tidningsspråk mycket heterogent, med stora skillnader mellan t.ex. ledarsidan och sportsidorna. Språket i kontrollkorpusen PRESS har i viss utsträckning renodlats genom att mer talspråkliga avdelningar inriktade mot barn och ungdom uteslutits. Korpusen omfattar ca 25 miljoner ord.

För att täcka det mer talspåksnära, vardagliga språkbruket kommer den kompletterande referenskorpusen UNGDOM att byggas upp. Denna korpus kommer att omfatta material ur de ovannämnda ungdomsavdelningarna ur tidningsmaterialet. Avsikten är naturligtvis att med denna korpus försöka täcka ett mer konkret språk, med just konkreta substantiv m.m.

Korpusarna måste kompletteras med lexikon, dels för etablering av lemman, dels för semantisk uppmärkning, så som beskrevs i avsnitt 2.2.2. För lemmatiseringen används EPOS, En Probabilistisk Ordklasstagger för Svenska (Johansson Kokkinakis 2003), vilken bygger på uppmärkning med den lexikaliska databasen SMDB (Svensk morfologisk databas), baserad på Svenska Akademiens ordlista (12:e uppl. 1998) och huvudsakligen innehållande samtliga möjliga böj-

ningsformer till ordlistans samtliga uppslagsord. För den semantisk uppmärkningen används en variant av databasen GLDB (Göteborgs universitets lexikaliska databas). Dessa databaser torde utgöra det fylligaste underlaget för ifrågavarande uppgifter, samtidigt som de är tillgängliga på institutionen och deras tekniska konstruktion är känd och dokumenterad. Databaserna beskrivs närmare i avsnitt 3.3.

Givet å ena sidan de antagna vokabulärkategorierna (A) – (D), å andra sidan textmaterialen i form av korpusarna OrdiL, PRESS och UNGDOM blir problemet hur man med hjälp av siffror för frekvens och spridning skall kunna utvinna vokabulärkategorierna ur korpusarna. En procedur för denna uppgift måste med nödvändighet vara automatisk, på grund av materialens storlek. Det kända verket *Nusvensk frekvensordbok 1 – 4 NFO*; 1970 – 1980) baserades på ett material på en miljon löpande ord och kostade c:a 40 manår att färdigställa med enbart manuell bearbetning. Därvid måste intuitiva begrepp som ”typiskt för viss text” uttolkas i termer av statistiska mått som frekvens och spridning. De principiella metoderna kommer att diskuteras i nästa avsnitt medan de tekniska detaljerna beskrivs i avsnitt 3.4.

2.4.3 Distributionella metoder

De enheter som det nedan talas om kan, enligt avsnitt 2.2.2, vara graford, s.k. homografkomponenter, lemman, lexem, flerordsenheter eller eventuellt ordfamiljer. I fortsättningen lämnas enheternas närmare karaktär därhän, eftersom den är föga väsentlig för de statistiska resonemangen.

Det grundläggande måttet för undersökningen är frekvensen för enheterna, ibland förtydligad till *absolut* frekvens. Därmed menas naturligtvis helt enkelt antalet förekomster av varje enhet i en given korpus. Eftersom korpusarna och delkorpusarna, t.ex. NO-delen av OrdiL, är olika stora måste man ofta räkna med *relativ* frekvens, lika med antalet förekomster av enheten, dividerat med korpusens storlek i löpande enheter, ord. Den relativa frekvensen är alltid ett litet tal; den högsta relativa frekvensen i svenska korpusar är ca 0,03 (3 %)

för ord som *och*. Vid redovisning av relativ frekvens kan det därför vara pedagogiskt att multiplicera den med ett lämpligt tal, så att man erhåller heltalsuppgifter, ca 30.000 *och* per miljon löpande ord, etc. Relativ frekvens diskuteras i detalj även i avsnitt 3.4.

Det för undersökningen viktigaste måttet avser enheternas spridning (dispersion) över delkorpusar. Ett spridningsmått för en viss enhet med avseende på en viss uppsättning delkorpusar anger i princip i vilken grad enhetens frekvens i delkorpusarna är proportionell mot delkopusarnas storlek eller, annorlunda uttryckt, huruvida enheten är jämnt fördelad över delkorpusarna. För OrdiL används ett spridningsmått som ursprungligen konstruerades av Juilland & Chang-Rodriguez 1964 och i modifierad form har använts i Nusvensk frekvensordbok. Måttet varierar mellan 1 (helt jämn spridning över delkorpusarna) och 0 (förekomst endast i en enda delkorpus) och är oberoende av delkorpusarnas storlek.

Som exempel kan nämnas listor i *NFO 2* över de 1.000 vanligaste lemmana per ämnessfär, korpusen är grovt uppdelad i sex ämnessfärer. I listan över vanligaste lemman i sfären ”naturvetenskap med tillämpningar”, s. 703 ff, finner man vid manuell excerpering bl.a. följande exempel, efter uteslutningar av namn, förkortningar och vissa särpräglade fall som *lucerngrönmjöl*, sorterade efter stigande spridningsmått och sekundärt efter stigande frekvens, se tabell 2.2.

Tabell 2.2 Lista över vanligaste lemman från NFO 2.

Lemma	Frekvens	Spridning
resistent	12	0,000
betning	14	0,000
utsäde	20	0,028
nederbörd	15	0,041
bakterie	23	0,049
oljeväxt	13	0,082
substans	19	0,082
cancer	28	0,093
preparat	16	0,106
organism	16	0,140
frö	28	0,150
gräsmatta	29	0,151
halt (subst.)	17	0,237
temperatur	38	0,263
karta	59	0,277
biologisk	24	0,310
hypotes	33	0,310

Som synes innehåller listan flera fackord från biologi och agronomi, jämte några tämligen allmänna men tydligen ämnestypiska ord som *gräsmatta* och några ”academic words” som *hypotes*. Vissa ord är lite överraskande som *temperatur* och *karta*. Med tanke på att ämnessfären i fråga är ganska bred och alls icke av lärobokskaraktär tycks resultatet av denna lilla undersökning tyda på att spridningsmåttet skulle kunna användas för att operationellt definiera projektets vokabulärkategorier. Resultatet indikerar också att en övre ”tröskel” för spridningsmåttet någonstans kring 0,3 skulle kunna användas för att operationellt definiera ämnestypiska ord. Det exakta tröskelvärdet får bestämmas efter preliminära undersökningar i korpusarna.

Formeln för det använda spridningsmåttet presenteras närmare i avsnitt 3.4.1. Måttet kan multipliceras med den absoluta frekvensen till ett nytt mått som kallas *modifierad* frekvens, vilken använts i *NFO*

för att etablera en s.k. basvokabulär. Modifierad frekvens tycks dock inte vara direkt aktuell för projektet.

Det står nu klart att det är möjligt att med lämpliga val av delkorpusar och tröskelvärden definiera de olika vokabulärkategorierna (A) – (C) ovan. För ett ”allmänspråkligt” ord fordras tydligen att det har tämligen jämn spridning över OrdiL och någon eller båda av referenskorpusarna. För ord i kategori (A) fordras dessutom tämligen hög relativ frekvens i UNGDOM. På motsvarande sätt fordras för ord i kategori (B) rimlig frekvens i OrdiL och PRESS, medan de kan få ha låg frekvens (och alltså dålig spridning) i UNGDOM. Ord i kategori (C) bör ha relativ frekvens över ett tröskelvärde i referenskorpusarna (för att garantera att de är ”allmänspråkliga”) men dålig spridning utanför ett visst ämne (för att garantera att det rör sig om ”ämnestypiska” ord). Den närmare metodiken återstår att utreda men uppenbarligen är det möjligt att utföra indelningen i vokabulärkategorier helt automatiskt.

Med de statistiska metoder som antytts ovan kan man inte utan vidare skilja mellan ämnestypiska ord och fackord. För att utskilja fackorden och facktermerna, kategori (D), kan man tänka sig att använda två olika metoder, var för sig eller i kombination. *Dels* kan man utnyttja det faktum att enheter med spridningsmått=0 enbart förekommer i en enda delkorpus. I tabell 2.2 över ämnessfären ”naturvetenskap med tillämpningar” kunde man då operationellt definiera de två första orden, *resistent* och *betning*, som ämnesspecifika. Denna procedur, som endast är en utbyggnad av den tidigare, kan åter förlöpa automatiskt. Den eventuella nackdelen är att metoden är rent statistisk och inte kan förhindra att tämligen allmänna enheter som slumpmässigt råkar förekomma i enbart någon lärobok, men inte i den stora tidningskorpusen, kommer att ”utnämnas” till fackord. *Dels* kan man, mer kvalitativt, använda sig av att information om enheternas typografi, textens layout o.d. har sparats. Då kan man antaga att enheter som i läroböckerna markerats särskilt, genom att kursiveras eller sättas i fetstil, placeras i rubrik eller inom speciella ramar, etc. normalt har en speciell karaktär och skulle kunna räknas som facktermer. Det är dock troligt att utnyttjande av informationen om typografi och s.k. texttyp fordrar manuella ingripanden.

2.5 Resultatredovisning

Resultaten av undersökningen av OrdiL-korpusen avses bli publicerade i en uppsättning listningar, antingen i elektroniskt format eller i form av en tryckt bok, då kanske som ett urval av de mer extensiva elektroniska listningarna. Sådan listningar kan innehålla en mängd uppgifter om enheterna: absolut och relativ frekvens, förekomst i olika delkorpusar, spridning i materialet, förekomst i andra typer av enheter gäller särskilt ettordsenheters förekomst i flerordsenheter, förekomst i särskilda texttyper, som rubriker etc., samt eventuella komplikationer som att vissa lemmaenheter kan representera olika lexem. Vidare kan listorna ordnas initialalfabetiskt men även finalalfabetiskt eller numeriskt efter fallande frekvens men kanske även efter stigande spridning, som i exemplet från 2.4.3. Det är därför självklart att de nedanstående listningarna endast skall betraktas som förslag och inte på något sätt kan täcka alla de nämnda dimensionerna. Som enda delkorpus har ämnet *fysik* använts men i en fullständig listning skulle på motsvarande sätt även de sju andra ämnena vara representerade. Kolumnen ”Annan text” skall i samtliga fall tolkas som sammanfattningen av rubriker, bildtext, typografiskt avvikande text etc.; d.v.s. allt som inte är normal, beskrivande s.k. brödtext.

2.5.1 Grafordsnivån

På grafordsnivån har varken separering av homografer eller sammanförande av böjningsformer ägt rum, varför sifferuppgifterna måste behandlas med reservation. Jfr också 2.2.2. Listorna på denna nivå är avsedda dels för jämförelser med andra, enklare frekvensundersökningar, dels för redovisning av flerordsenheter. I listorna har *homografa* enheter markerats med asterisk. Relativ frekvens har multiplicerats med 1.000.000 och avrundats till heltal.

Alfabetisk listning av graford

Uppgifterna om förekomst i de olika kategorierna redovisas här med absolut frekvens. Observera att siffran för absolut och relativ frekvens totalt avser hela materialet; kanske ca 26 miljoner ord.

Tabell 2.3 Exempel på alfabetisk listning av graford.

Enhet	Rel. frekv.	Abs. frekv.	OrdiL	Flerordsenh.	Annan text	Fys.
hamnar*	53	2154	87	0	4	43
lade	49	2786	51	14	0	7
märke	31	1491	37	35	0	18

Alfabetisk listning av flerordsenheter

Tabell 2.4 Exempel på alfabetisk listning av flerordsenheter.

Enhet	Rel. frekv.	Abs. frekv.	OrdiL	Annan text	Fys.
lade märke till	2	104	13	0	6
lägga märke till	7	348	21	0	11

Ämnestypiska och fackordsliknande graford i fysik efter stigande spridningsmått

De graford som har spridningsmått 0,000 kan säkert betraktas som ett slags fackord. I övrigt kan t.ex. förekomst i ”Annan text” få avgöra till vilken kategori enheterna bör räknas.

Tabell 2.5 Exempel på ämnestypiska och fackordsliknande graford i fysik.

Enhet	Rel. frekv.	Abs. frekv.	OrdiL	Annan text	Fys.
elektron subst. -en -ar	2	145	37	6	29
hamn subst. -en -ar	112	5613	46	2	0
hamna verb -ade -at	81	4102	73	1	67
lägga* verb -ade -t	447	23017	483	24	63
trycka* verb -te -t	74	3762	103	2	64

2.5.2 Lemmanivån

Listorna på denna nivå är jämförbara med ordlistor, t.ex. SAOL. Därigenom ger de en säkrare bild av ordförrådet än listorna på grafordsnivå, åtminstone för längre lemman som inte är polysema. Varje enhet sammanfattar i princip samtliga böjningsformer, vilka eventuellt kunde anges i listorna grupperade under sin representationsform. I listorna har *polysema* enheter markerats med asterisk. Troligen måste listorna inskränkas till huvudordklasserna substantiv, adjektiv och verb, eftersom lemmatisering (inklusive homografseparering) för övriga ordklasser är onödigt betungande, utan att tillföra särskilt intressant information.

Alfabetisk listning av lemman

Tabell 2.6 Exempel på alfabetisk listning av lemman.

Enhet	Rel. frekv.	Abs. frekv.	OrdiL	Annan text	Fys.
elektron subst. -en -ar	2	145	37	6	29
hamn subst. -en -ar	112	5613	46	2	0
hamna verb -ade -at	81	4102	73	1	67
lägga* verb -ade -t	447	23017	483	24	63
trycka* verb -te -t	74	3762	103	2	64

Alfabetisk listning av vissa flerordslemman

Tabell 2.7 Exempel på alfabetisk listning av flerordslemman.

Enhet	Rel. frekv	Abs. frekv.	OrdiL	Annan text	Fys.
lägga märke till	14	682	43	0	20

Tabell 2.8 Exempel på ämnestypiska och fackordliknande lemman i fysik efter stigande spridningsmått..

Enhet	Spridn.	Rel. frekv.	Abs. frekv.	OrdiL	Annan text
primärspole subst.	0,000	0	13	13	2
partikelmodell subst.	0,000	0	14	13	0
elektron subst.	0,030	3	162	46	8

2.5.3 Lexemnivån

På denna nivå är listorna jämförbara med definitionsordböcker och är alltså de som bäst beskriver de innehållsliga aspekterna av ordförrådet i läroböckerna. Tyvärr måste listningarna på lexemnivå inskränkas till de mest frekventa lemmana ur huvudordklasserna, med en frekvens i OrdiL större än 4. Frekvenssiffrorna nedan avser uteslutande OrdiL-materialet. Huvudlistan nedan kan betraktas som ett komplement till lemmalistan ovan. I listan har lexem med utvidgade eller överförda betydelser, *underbetydelser*, markerats med asterisk.

Alfabetisk listning av lexem

Tabell 2.9 Exempel på alfabetisk listning av lexem.

Enhet	OrdiL	Annan text	Fys.
trycka verb /1*	72	2	68
trycka verb /3	31	0	0

En lista över ämnestypiska lexem och facklexem skulle egentligen komplettera framställningen. Med ledning av föregående torde det vara lätt att i princip rekonstruera en sådan lista, varför den utelämnas här.

2.5.4 Kompletterande redovisningar

Många typer av kompletteringar skulle kunna tänkas. Här skall två möjliga listor nämnas: dels en lista över namn- och termliknade flerordsenheter, dels en ”ordförklaring” till lexemnumren från lexemlistan ovan. Listorna kan i princip ha följande uppläggning:

Flerordnamn och flerordstermer

Tabell 2.10 Exempel på lista av namn- och termliknande flerordsenheter.

Enhet	Förklaring
med blotta ögat	utan optiska hjälpmedel (som kikare eller mikroskop)
de magdeburgska halvkloten	en konstruktion för att bevisa lufttryckets styrka (på 1600-talet)

Betydelsebeskrivningar av lexem

Tabell 2.11 Exempel på lista av betydelsebeskrivningar av lexem.

Enhet	Definition
trycka verb /1	’påverka genom direkt kontakt (i riktning från kraftkällan)’
trycka verb /3	’mångfaldiga (skrift) i tryckpress’

Referenser

Allén, S. 1970. *Nusvensk frekvensordbok 1 baserad på tidningstext*. Data Linguistica 1.Stockholm. Förkortad *NFO 1*

Allén, S. 1971. *Nusvensk frekvensordbok 2 baserad på tidningstext*. Data Linguistica 2. Stockholm. Förkortad *NFO 2*

Berg, S. 1978. *Olika lika ord. Svenskt homograflexikon*. Data Linguistica 12. Stockholm.

Johansson Kokkinakis, S. 2003. *En studie över påverkande faktorer i ordklasstaggning. Baserad på taggning av svensk text med EPOS*. Avhandling, Humanistiska fakulteten, Göteborgs universitet.

Järborg, J. 1989. *Betydelseanalys och betydelsebeskrivning i Lexikalisk databas*. Språkdata, Göteborgs universitet. Dupl.

Järborg, J. 2003a. *Formaliserade semantiska samband mellan enheter i GLDB*. Research Reports from the Department of Swedish, Göteborg University. GU-ISS-03-1.

Järborg, J. 2003b. *Semantisk uppmärkning. Metoder, problem och resultat*. Research Reports From the Department of Swedish, Göteborg University. GU-ISS-03-2.

Järborg, J. 2005. *Introduktion i datamaskinell lexikologi*. Språkdata, Göteborgs universitet.

Kokkinakis, D., Järborg, J. & Cederholm, Y. 2001. ”Lexical and Textual Resources for Sense Recognition and Description”. I *Proceedings of the 3rd LREC*. Las Palmas.

Nation, P. 2001. *Learning vocabulary in another language*. Cambridge Cambridge University Press.

Nationalencyklopediens Ordbok, 1995-96. Höganäs.

Qian, D. 2002. Investigating the relationship between vocabulary knowledge and academic reading performance: an assessment perspective. *Language Learning* 52.

Rosch, 1977. ”Human Categorization”. I N. Warren (Ed.) *Studies in Cross-Cultural Psychology*, vol. 1. New York.

de Saussure, F. 1916. *Cours de Linguistic Général*. Lausanne & Paris

Stroud, C. 1979. Kontrastiv lexikologi. I Hyltenstam, K. (utg.), *Svenska i invandrarperspektiv. Kontrastiv analys och språktypologi*. Lund: Liber läromedel.

Svensk ordbok, 1986. Stockholm.
Svenska Akademiens ordlista, 1998. 12:e upplagan. Stockholm. Förkortad *SAOL*.
Viberg, Å. 2004. Lexikal utveckling i ett andraspråk I K. Hyltenstam & I. Lindberg (red.) *Svenska som andraspråk i forskning, undervisning och samhälle*. Lund: Studentlitteratur.

3. Språkteknologiskt arbete i OrdiL-projektet

Sofie Johansson Kokkinakis

Språkteknologiskt arbete som har utförts inom OrdiL-projektet har under 2004 gjorts av Dimitrios Kokkinakis (DK) samt under 2003, 2005-2006 av Sofie Johansson Kokkinakis (SJK). Avsnitt i texten där DK har utfört arbete markeras med initialer, övrig text är skriven av SJK.

Inledningsvis diskuteras utgångspunkt och frågeställningar om material, avsnitt 3.1. Därefter görs en relativt detaljerad beskrivning av bearbetningen av material, avsnitt 3.2. I avsnitt 3.3 redovisas de tillämpningar som skapats, såsom databaser, gränssnitt för semi-automatisk semantisk disambiguering, Göteborgs Lexikaliska Databas, webb-baserat gränssnitt för sökning innehållande konkordanser och HTML-versioner av läromedelstexter. I avsnitt 3.4 ges information om statistisk information i läromedelstexterna som skapats inom projektet, och även olika former av tabeller med statistiska data samt användning av ett program som redovisar frekvensintervall i texter.

3.1 Material

I detta avsnitt ges en bakgrundsbeskrivning av projektets materiella utgångsläge samt en diskussion om hur man bearbetar givet material för att uppnå målen i projektet.

3.1.1 Om material

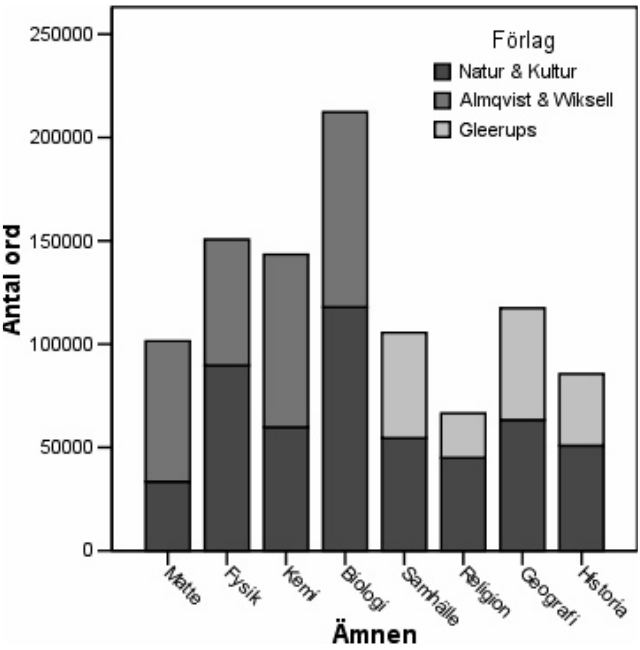
De läromedelstexter som beskrivs i projektet kommer ifrån tre olika förlag; Natur & Kultur, Almqvist & Wiksell och Gleerups. Ämnen som representeras i läromedelsmaterialet är: fysik, kemi, biologi (NO-ämnen) och matematik. Dessutom finns samhällskunskap, historia, religion och geografi (SO-ämnen). Sammanlagda antalet ord är 983.007 vilket fördelar sig på följande sätt över ämnena, SO - 375.112 och NO - 607.895. I figur 3.1 visas fördelningen av ord över ämnen samt förlag.

3.1.2 Anknytning till projektets mål

Ett av projektets mål är att alla texter ska sammanfogas till en större text, en s.k. *korpus*[5] för att man ska kunna undersöka språket i läromedelstexter som en helhet. Korpusar kan användas för att påvisa skriftliga belägg för teoretiska antaganden.

Ett annat mål är att utifrån läromedelskorpusen generera ordlistor sorterade dels på ämne men också efter frekvens eller bedömd svårighetsgrad. Speciellt svåra sådana ord kan vara vissa tvetydiga ord, fasta uttryck samt ovanliga fackord.

[5] En korpus är en textsamling eller en textmängd. Se definition: **CORPUS** (13c: from Latin corpus body. The plural is usually corpora) (1) A collection of texts, especially if complete and self-contained: the corpus of Anglo- Saxon verse. (2) Plural also corpuses. In linguistics and lexicography, a body of texts, utterances or other specimens considered more or less representative of a language, and usually stored as an electronic database. The Oxford Companion to the English Language. Oxford:265-266)



Figur 3.1 Fördelning av ord över ämnen samt förlag.

3.1.3 Vilken typ av information är intressant?

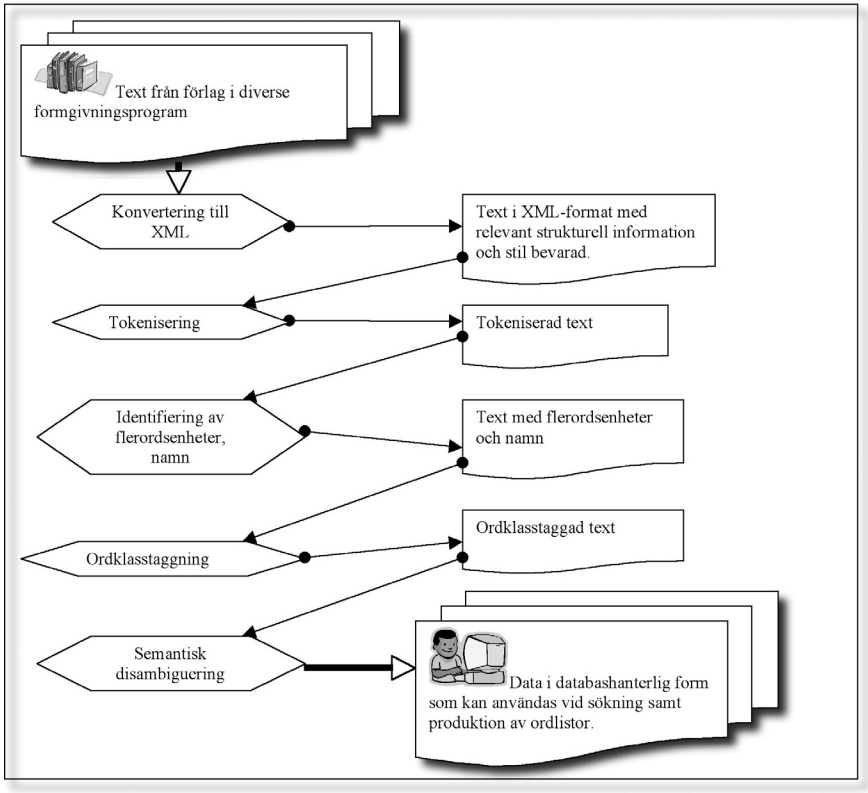
Vid projektets start bestämdes att en viss textspecifik information skulle bevaras. Detta gjordes med utgångspunkt från att ord och uttryck har olika utseende beroende på var i texten de förekommer. En rubrik är t.ex. ofta kort och utan markerat meningsslut. En bildtext eller text i en bild kan däremot bestå av fackspråkliga uttryck medan texten i en tabell ofta innehåller sifferuttryck i kombination med termer. För att ha möjligheten att dra slutsatser om innehåll och språklig information om ord i olika texttypiska uttryck beslutades att följande information skulle bevaras: kapitelrubrik, underrubrik, bildtext, text i bild, tabellrubrik, tabelltext, rutrubrik, ruttext, faktarubrik, faktatext och brödtext (normal text). Vidare sparades också uppgifter om citat, webb-länk och titel. Av stiltyper sparas information om fetstil, kursiv stil, understruken text och normal stil.

3.2 Bearbetning

I detta avsnitt beskrivs varför en text behöver bearbetas, på vilket sätt och på vilken nivå bearbetningen bör ske samt vad som krävs för att uppnå projektets mål.

3.2.1 Vad behövs och varför?

För att kunna använda korpusen med läromedelstexter i ett sökprogram anpassat för sökningar och frågor som ställs av forskare samt generering av ordlistor behöver den genomgå flera analyser och bearbetas ytterligare.



Figur 3.2 Bearbetningssteg av läromedelsmaterial i OrdiL-korpusen

Alla bearbetningar är anpassade för att slutmålet ska kunna uppnås. Ett av projektets mål är dels att kunna generera relevanta ordlistor men också att kunna söka i en text med grammatisk information och betydelsebeskrivningar för de flesta ord. Figur 3.2 illustrerar de olika processerna en text som denna behöver gå igenom.

De nämnda bearbetningsstegen beskrivs mer ingående i avsnitt 3.2.2-3.2.6.

3.2.2 Konvertering till XML

Läromedelstexterna levererades från förlagen dels i tryckt bokformat och dels i datafiler i de formgivningsprogram som respektive förlag använde. Formgivningsprogrammen var Quark-XPress (internt QuarkXPress-format) och Adobe Acrobat (pdf-format). Bearbetning av fyra av läromedlen har gjorts av DK.

För att kunna göra materialet och korpusen sökbar i datateknisk mening måste texterna överföras till ett format som kan användas vid bearbetning och är läsbart för bearbetningsprogram. Ett sådant format är i vårt fall ett rent textformat med strukturell och betydelsebärande information bevarad. XML (*Extensible Markup Language*) är ett format som används bl.a. för att beskriva strukturell information i en text. Följande beskrivning av XML finns på World Wide Web consortium.

> ”The Extensible Markup Language (XML) is a subset of SGML (Standard Generalized Markup Language)... Its goal is to enable generic SGML to be served, received, and processed on the Web in the way that is now possible with HTML. XML has been designed for ease of implementation and for interoperability with both SGML and HTML.” (http://www3.org/XML).

En XML-uppmärkning av en rubrik kan se ut på följande sätt:

```
<underrubrik>Handel</underrubrik>
```

Strukturell information kan t.ex. beskriva om en del av en text är en rubrik, brödtext, eller bildtext. Betydelsebärande information är sådant som inte omfattas av strukturell information. Det kan vara uppgifter om en texts egenskaper, t.ex. citat, webb-länk, titel eller liknande där text har ett speciellt utseende vilket markeras med fetstil, kursiv stil eller som understruken text. Denna information är inte att förväxla med vanlig stilinformation som har med textens formatering att göra. Det kan vara markering av s.k. svåra ord eller ord av speciell betydelse. För att kunna urskilja s.k. svåra ord i t.ex. kursiv stil från ord som förekommer i citat där ett större stycke är kursiverat krävs en särskiljning mellan de två. Annan information som bevaras är om var ett stycke startar och slutar.

3.2.3 Tokenisering

Ett andra steg i bearbetningsprocessen är *tokenisering*. Tokenisering (eng. tokens) innebär att ord i en text separeras från varandra för att möjliggöra vidare bearbetning av enskilda ord. Det handlar om att urskilja vad som är ett ord, en förkortning eller en sammanhängande enhet. Man separerar skiljetecken från ord i de fall då de ska separeras. Ett exempel på detta är:

Han gick hem.

En tokenisering av satsen blir då:

Han gick hem .

Ett annat exempel är:

Hon är vegetarian men gillar ändå kött, t.ex. pizza och hamburgare.

vilket efter tokenisering får motsvarigheten:

Hon är vegetarian men gillar ändå kött , t.ex. pizza och hamburgare .

Notera att ”t.ex.” inte separeras efter skiljetecken eftersom det är en sammanhängande enhet som ska tolkas som den är.

Andra skiljetecken som oftast inte förekommer tillsammans med ord är:

; : , ! ? ( ) [ ] { } - /

Men det finns undantag, som t.ex. *2-3* och *TV:n*. Sådana undantag måste definieras i förväg för att tokeniseringen ska bli korrekt. En enhet som *2-3* kan inte representeras som en undantagsregel som den är utan måste göras om till ett s.k. *reguljärt uttryck*. Ett sådant uttryck används för att skapa mer generella regler. Då anges istället *[0-9]+-[0-9]+*, 'numerärt uttryck-numerärt uttryck'. Ett plus betyder i sammanhanget en eller flera förekomster av något, t.ex. en siffra. Reguljära uttryck kan oftast användas i sökningar i webb-verktyg, t.ex. Google eller sökverktyget i OrdiL-projektet. Antag t.ex. att man är intresserad av numerära uttryck som föregås av textsträngen *mer än* så formuleras sökningen på följande sätt: *mer än [0-9]+*. Man får då träffar som: *mer än 250 år sedan*, *mer än 80 %* etc.

Det är dock aldrig trivialt att tokenisera en text då den ofta präglas av ämnet. Skillnaden är stor mellan en text i matematik (figur 3.3) och en text i historia (figur 3.4) eftersom alla matematiska formler och uttryck ofta innebär undantag till reguljära tokeniseringsregler medan en text i historia ofta består av löpande text eventuellt med inslag av citat.

Cirkelns area = π × radien × radien

A = π × r × r = π × r2

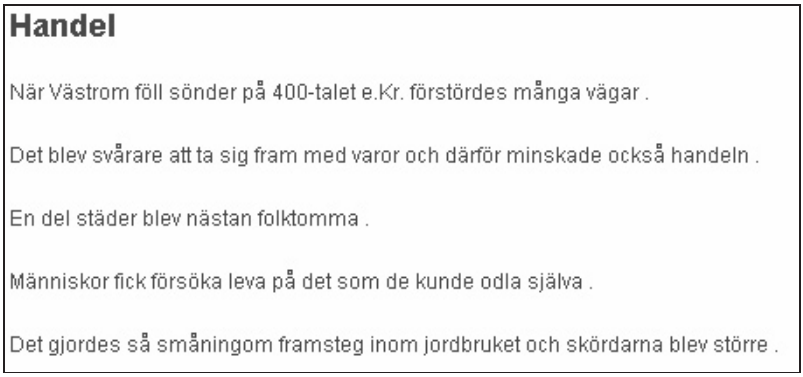
π ≈ 3,14

r2 = " r upphöjt till 2 " är ett annat sätt att skriva *r* × *r*

52 = 5 upphöjt till 2 betyder 5 × 5 , alltså 25

Figur 3.3 Exempel på en text i ämnet matematik.

I figur 3.3 ges exempel på svårigheter att förutse matematiska formler. Då separeras alla delar, t.ex. π × *radien* × *radien*. Det är också önskvärt att *3,14* inte tolkas som vanlig text vilken i normalfallet separeras vid kommatecken. Här behövs därför speciella regler för tolkning av numeriska uttryck.



Figur 3.4 Exempel på en text i ämnet historia.

I figur 3.4 finns exempel på en annan typ av text, nämligen en historisk. Där upptäcker man ändå att det finns tidsuttryck *400-talet* och förkortningar *e.Kr.* vilka måste tolkas som enheter som inte ska splittras.

3.2.4 Flerordsenheter och namn

För att kunna föra vissa diskussioner om det som kommer näst i bearbetningen av korpusmaterialet är det nödvändigt att gå igenom vad begreppet ord innebär. Därför används SAG (Svenska Akademiens Grammatik 1999) för vissa definitioner medan vi i projektet har en mer generell syn på lexikaliska termer.

> *Lexikonord* är orden i talarens ordförråd och består av rotmorfem eller av flera morfem som är fast sammanfogade till lexikonord genom sammansättning eller avledning, t.ex. ”smal, långsmal, smalna”.
>
> *Lexikaliserad ordförbindelse* är ett mellanting mellan ett flermorfemigt lexikonord och en syntaktiskt konstruerad ordgrupp och har en enhetlig betydelse vilken inte utan vidare kan föutsägas utifrån betydelsen hos de enskilda ord som ingår i förbindelsen, t.ex. ”ge upp”.

Det finns också två grupper av lexikaliserade ordförbindelser, nämligen *fixerade enheter* (där de ingående enheterna är omöjliga att skilja åt, t.ex. över huvud taget/överhuvudtaget), och *ej fixerade enheter* (i vilket de ingående delarna kan delas upp och böjas, t.ex. ”ger inte upp”, ”ångrade han sig”). Fixerade enheter motsvarar vad vi i projektet benämner *flerordsenheter*, se också Kokkinakis (2001:69-72).

Inom Språkteknologi definieras *word token* som en teckensekvens, där en sekvens omges av mellanrum och det kan jämföras med svenskans *textord*. Brukliga beteckningar inom Språkteknologi och Korpuslingvistik definieras av Leech (Garside et al. 1997). *Ortografiska ord* (textord) anges som avgränsade enheter i skriven text vilka föregås och följs av mellanslag. Dessa skiljer sig från *morfosyntaktiska enheter* (syntaktiska ord) vilka han definierar som de textord (word tokens) vilka man identifierar vid grammatisk taggning. Se vidare om grammatisk taggning i avsnitt 3.2.5. Dessa två begrepp skiljer sig åt på tre punkter:

1) *Flerordsenheter* (multiwords) - mer än ett ortografiskt ord motsvaras av ett annat morfosyntaktiskt ord. T.ex. ”in spite of” vilket kan tolkas som tre ortografiska ord, men de taggas med fördel som en morfosyntaktisk enhet.

2) *Sammanslagningar* (mergers) - där ett ortografiskt ord motsvarar mer än ett annat ortografiskt ord. Exempel på sådana kan vara proklitiska former ”t’” i ”je t’aime” eller enklitiska former ”n’t” i ”hasn’t”. De skrivs som en ortografisk enhet vilket också tolkas som en morfosyntaktisk enhet.

3) *Sammansättningar* (compounds) - kan tolkas så att beroende på analysen motsvarar en eller flera ortografiska ord en eller flera morfosyntaktiska ord. Exempel på en typisk sammansättning är ”rainbow” där ordet kan delas upp i delar där delarna i sig är fristående ord. Det finns dock undantag till detta i form av s.k. skenord, där ord som ”York-San” (från New York-San Francisco flights”, vilket faktiskt kan uppstå då ortografiska ord ska motsvaras av morfosyntaktiska ord.

I detta tredje steg i bearbetningen av en text identifieras ortografiska sammanhängande flerordsenheter med motsvarande morfosyntaktiska sådana. Fördelen med att genomföra omvandling av ortografiska ord till motsvarande morfosyntaktiska enheter i ett separat steg är att det även ska göras vid ordklasstaggning eller morfosyntaktisk uppmärkning, se avsnitt 3.2.5. Problemet med att utföra en sådan identifiering i samband med ordklasstaggningen är att det finns flera olika tillämpbara regler för att avgöra vilken ordklass en enhet ska få. Risken ökar då att en felaktig tolkning väljs, så därför underlättas taggningen av att existerande morfosyntaktiska enheter redan har identifierats.

Identifiering av namnuttryck sker i anslutning till identifiering av flerordsenheter. Till detta används ett speciellt datorprogram speciellt utvecklat för namnigenkänning, (Kokkinakis 2004), som dels innehåller namnlistor, de flesta som reguljära uttryck och dels andra regler som kan härleda namnuttryck ifrån vissa kombinationer av versaler etc. De flerordsuttryck som identifierats innehåller alltså till viss del namnuttryck. Då markeras namnuttryck såsom *Kalle Persson* som en morfosyntaktisk enhet även om det är två skilda ortografiska ord.

3.2.5 Ordklasstaggning och disambiguering

Ordklasstaggning innebär att ord utifrån en given tagguppsättning får en grammatisk analys och markeras med en *etikett/tagg* (eng. tag), t.ex. *verb*, *adjektiv* etc. En sådan tagg kan innehålla olika mycket grammatisk information. Det enklaste är att endast markera att ett ord tillhör en viss ordklass, substantiv, verb, adjektiv etc. och inget mer. Men vill man vara mer specifik kan även tempus anges för verb, eller så kan för substantiv anges numerus, bestämdhet, genus etc. Vad som ska uppmärkas anges i en tagguppsättning där alla möjliga taggar ingår. Man brukar tala om storleken på en tagguppsättning. Storleken anger hur många kombinationer som är möjliga för t.ex. substantiv. Antag att det finns åtta möjliga taggar:

substantiv utrum singularis bestämd	(stolen)
substantiv neutrum singularis bestämd	(bordet)
substantiv utrum pluralis bestämd	(stolarna)
substantiv neutrum pluralis bestäm	(borden)
substantiv utrum singularis obestämd	(stol)
substantiv neutrum singularis obestämd	(bord)
substantiv utrum pluralis obestämd	(stolar)
substantiv neutrum pluralis obestämd	(bord)

I exemplet anges *bord* på två ställen. Det är i detta fall fråga om intern homografi.

Men ordklasstaggning innebär inte att man endast anger en tagg per ord utan i hälften av fallen är ett ord *homograft*[5] vilket innebär att det kan tolkas grammatiskt på mer än ett sätt. Vidare finns två relevanta underordnade begrepp: *extern homografi*[6] och *intern homografi*[7].

För att ytterligare klargöra skillnaden mellan extern och intern homografi visas tabellerna (tabell 3.1 & 3.2). Se också avsnitt 2.2.1 i denna rapport.

Tabell 3.1 Exempel på homografseparering och lemmatisering (NFO, Allén 1970).

C	stack		sticka		sticka	sticka					
	nn –en		vb –ø		nn -n	vb –ad					
B	stacken	stack	stack	sticker	sticka	sticka	sticka	sticka	stickar	stickat	stickat
		nn –en	vb -ø		vb -ø	nn -n	vb -ad inf	vb –ad inf		vb -ad sum	vb -ad ptp
A	stacken	stack		sticker	sticka				stickar	stickat	
	1	2		3	4				5	6	

Tabell 3.2 Fördelning av graford, homografkomponenter och lemman från tabell 3.1.

A	grafordsnivå
B	homografkomponentnivå
C	lemmanivå
1,3,5	heterografi
2	extern homografi
4	extern och intern homografi
6	intern homografi

[5] Homografi är ”... ord som till sin skriftform är identiskt med ett annat ord men olika beträffande ljudform, ursprung och betydelse...” Berg (1970)
[6] Extern homografi är ”... den relation som råder mellan i skrift identiska former tillhörande olika lemman.” Allén (1970).
[7] Intern homografi är ”...den relation som råder mellan i skrift identiska reflexionsformen (grundformen inräknad) inom ett och samma lemma” Allén (1970)

För att lösa problemet med homografi använder man sig av *disambiguering*. Det innebär att olika ledtrådar om ordets kontext används för att avgöra vilken ordklasstagg som är den korrekta i sammanhanget. Detta kan göras för hand, men blir tidsödande vid större datamängder. Därför används olika metoder i datorprogram för att utföra disambiguering automatiskt. Observera att det idag inte finns någon metod som med 100 % korrekt resultat kan utföra automatisk disambiguering. Normalfallet ligger mellan 93-98 % korrekt disambiguerad text. Det finns givetvis flera påverkande faktorer i sammanhanget såsom att stigande storlek på tagguppsättningen påverkar resultatet negativt liksom att en text med många homografa ord ökar svårigheten vid disambiguering. Dessutom beror ordklasstaggningen på vilket lexikon man använder sig av och hur många homografa former det innehåller för ett ord (se vidare Johansson Kokkinakis 2003).

Denna typ av uppmärkning då man förser ett ord med morfosyntaktisk information brukar ske mer eller mindre automatiskt med hjälp av datorprogram. De datorprogram som används brukar använda sig av olika metoder och tekniker. Det finns dels lingvistiska och dels datadrivna tekniker. Eftersom de tekniker som används i detta projekt är datadrivna kommer en kort introduktion här att ges om dem. Datadrivna tekniker går ut på att *maskininlärning* används för att förbättra ett datorprogram ytterligare. Daelemans (1999) beskriver detta som en underdisciplin till artificiell intelligens. Förbättringstekniker går ut på att antingen lära programmet genom erfarenhet eller genom att omstrukturera redan insamlad information.

Det finns också olika datatekniska metoder. Några av dessa är N-gram, Hidden Markov-modeller, neurala nätverk/konnektionism, fallbaserade metoder, memory-based learning, transformationbased learning och maximum entropy. Se vidare i Johansson Kokkinakis (2003). De metoder som används i projektet är N-gram genom verktyget EPOS[8] (En Probabilistisk Ordklasstagger för Svenska)

[8] EPOS, En Probabilistisk Ordklasstagger för Svenska, Johansson Kokkinakis (2003). Ett disambigueringsverktyg utvecklat för att disambiguera texter annoterade med SMDB (Svensk Morfologisk DataBas) vilken bygger på Svenska Akademiens Ordlista.

använts för att ordklasstagga texterna i ett första skede och därefter en TnT-tagger (Brants 2005) tränad på SUC (Stockholm-Umeå-Corpus) (Källgren 1998). N-gram är en metod som använder sig av kedjor av grammatiska komponenter eller ortografiska ord samt sannolikhetsmått för att avgöra vilken kombination som är mest trolig då det finns två eller flera möjliga val. En Hidden Markov-modell bygger på samma princip som N-gram, men istället för att ange alternativen som säkra möjligheter representeras de som icke-synliga regler.

3.2.6 Semantisk disambiguering och lemmatisering

Vi har tidigare talat om syntaktisk disambiguering, men det finns även semantisk disambiguering. De båda skiljer sig genom att en syntaktisk enhet inte enbart kan tolkas på ett sätt semantiskt sett. För att säkerställa kvaliteten på disambigueringen av ordklasstaggningen har Järborg i projektet även utfört manuell semantisk disambiguering i ca 25 % av korpustexten. Detta gäller framför allt ämnesord vilka mestadels är substantiv och verb. I den semantiska disambigueringen förs olika semantiska betydelsebeskrivningar till ett graford. Beroende på kontext får graforden olika beskrivningar. Semantisk disambiguering gör syntaktisk disambiguering överflödig i och med att då ett lexem av en ordform identifieras så kan det inte förekomma någon syntaktisk homografi. Men tvärtom gäller inte. Se vidare i avsnitt 2.2.2 i denna rapport. En mer datateknisk beskrivning av semantiskt disambiguering ges också i avsnitt 3.3.2.

3.3 Tillämpningar

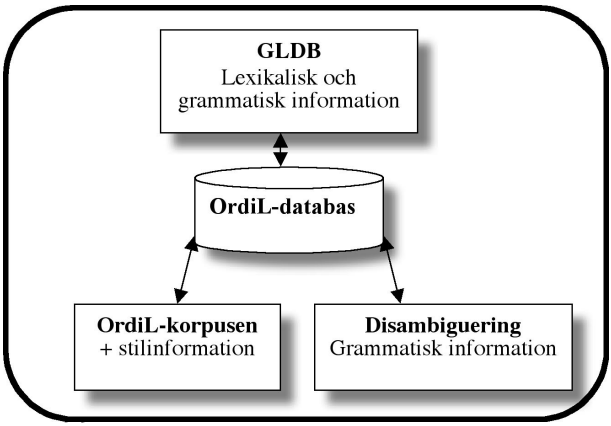
Den datorbaserade tillämpning som har skapats inom projektet kan sägas bestå av fyra delar. De är en databas, ett webb-baserat gränssnitt för semi-automatisk semantisk disambiguering i korpusen, ett webb-baserat gränssnitt för sökning i läromedelskorpusen samt sammanställningar av statistiska data i korpusen.

3.3.1 Databas

Databasen i OrdiL består av flera delar. Den innehåller ett lexikon (GLDB), en representation av texten i OrdiL-korpusen med tillhörande strukturell information och stilinformation samt grammatisk information för den del av korpusen som är ordklasstaggad och partiellt disambiguerad, se figur 3.5.

GLDB - lexikondelen

För att få tillgång till ett information om ett lemma, ordklass, lexem och definitioner används ett maskinläsbart lexikon GLDB (Göteborgs Lexikaliska Databas, Järborg 2003). Detta lexikon används främst i samband med semantisk disambiguering då lemma- och lexemnummer som anges för alla ord i läromedelskorpusen är direkt kopplade till lexikonet och dess definitioner. Se även avsnitt 2.4 i denna rapport.



Figur 3.5 De olika delarna i OrdiL-databasen

Då uppslag i GLDB av en grundform *glas* anges, som i figur 3.6, kan information om olika lemman och lexem erhållas som i figur 3.7.

Figur 3.6 Gränssnitt för uppslag av grundform och hämtning av lexikoninformation.

hittade: **3** lemma som innehåller **glas**

LEMMA	OKL	LMLX	DEFINITION
glas	subst.	1/1	ett hårt, glänsande, genomskinligt ämne
glas	subst.	1/2	dryckeskärl av glas
glas	subst.	1/3	halvtimma

Figur 3.7 Information i GLDB om *glas*.

Databasversion av OrdiL-korpusen

En databasversion av en textkorpus innebär att innehållet i texten förblir detsamma men att formatet för hur texten lagras skiljer sig. Databasen i det här fallet är skapad i MySQL och innehåller tabeller för all tänkbar information. Anledningen till att man använder sig av databaser i detta sammanhang är dels för att det är praktiskt och utrymmesbesparande, men det är också en stor fördel då man avser att söka i materialet t.ex. i ett webb-baserat gränssnitt.

Då texten representeras i tabellformat görs ordsegmentering radvis för att varje graford eller som diskuterats tidigare morfosyntaktisk enhet ska kunna betraktas separat. Varje enhet får ett id-nummer. Vidare kopplas detta id-nummer till annan typ av information som associeras

med ordet. I tabell 3.3 ges ett exempel på hur texten kan se ut när den är segmenterad och i databasformat.

Tabell 3.3 Beskrivning av strukturell information och stilinformation i databasformat.

Id-nummer	Graford	Texttyp	Stil	Källa
1	Livet	underrubrik	normal	biologibok1
2	börjar	underrubrik	normal	biologibok1
3	med	underrubrik	normal	biologibok1
4	en	underrubrik	normal	biologibok1
5	cell	underrubrik	normal	biologibok1
6	Alla	brödtext	normal	biologibok1
7	levande	brödtext	normal	biologibok1
8	organismer	brödtext	normal	biologibok1
9	som	brödtext	normal	biologibok1

Av exemplet framgår också att det finns strukturell information om texttyp och stilinformation som gäller formatering. De båda informationstyperna bevaras och lagras i databasformat och kopplas till respektive id-nummer som varje graford representerar i den löpande texten.

Antag att man har en mening som ”<underrubrik>Livet börjar med en cell</underrubrik>”, fortsättningen: ”<brödtext>Alla levande organismer som har beskrivits i det här kapitlet... </brödtext>”. Meningarna skulle i databasformat beskrivas som i tabell 3.3.

Fördelen med att spara information om stil och struktur är att det senare kommer att vara möjligt att göra sökningar på text enbart med sökvillkoret att den ska ha texttypen *underrubrik*, eller söka på *normal* eller *fet* stil.

Den grammatiska informationen som har tillfogats korpusen läggs också till i databasen. Först genomgår texten en ordklasstaggning där varje form av ett graford också har ett lemma/lexem och en ordklass. Eftersom arbetet med disambigueringen av korpusen sker efter taggningen och stegvis, så associeras en del graford med mer än en ordform/lemma/ordklasstagg medan andra endast har en. Information om och när disambiguering har ägt rum sparas. Ett exempel på en fras som är lemmatiserad, ordklasstaggad och uppmärkt med lemma- och lexemnummer illustreras i tabell 3.4.

Tabell 3.4 Exempel från databasen med uppgifter som rör ordklasstaggning och semantisk disambiguering.

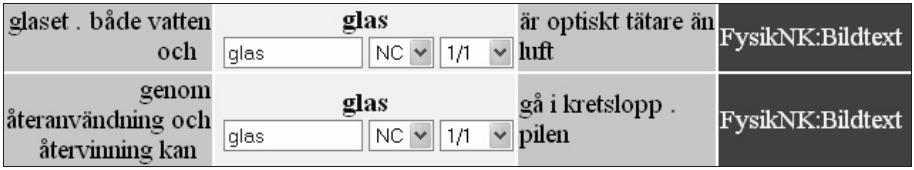
Id-nr	Graford	Lemma	Pos	Lm/Lx	Sms	Källa
1	Lådor	låda	NCUPNI	1/1	no	fysikbok1
2	och	och	CC	1/1	no	fysikbok1
3	liknande	likna	AP0**N*	1/1	no	fysikbok1
4	föremål	föremål	NCNPNI	1/1	no	fysikbok1
5	med	med	S	1/4	no	fysikbok1
6	räta	rät	AQP*PN*	1/2	no	fysikbok1
7	vinklar	vinkel	NCUPNI	1/1	no	fysikbok1
8	(	(	FC	1/1	no	fysikbok1
9	”	”	FC	1/1	no	fysikbok1

3.3.2 Webb-baserat gränssnitt för semi-automatisk semantisk disambiguering

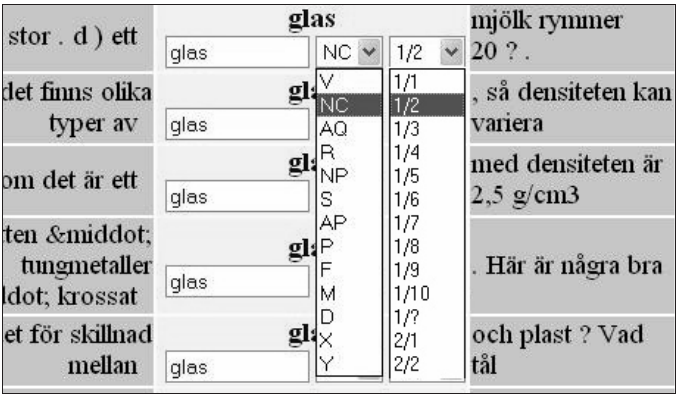
I detta avsnitt ges en beskrivning av den datatekniska delen av semantisk disambiguering av läromedelstexterna. För mer information om arbetet och tillvägagångssätt hänvisas till avsnitt 2.2.2 i denna rapport. I detta avsnitt beskrivs arbete utfört av DK såsom implementering av gränssnitt och databas för semantisk disambiguering.

I ett första skede bestämdes att lemmanummer och lexemnummer skulle bevaras och även ytlig grammatisk information i form av initi-

ala ordklasstaggar. Lemma- och lexemnummer anger vilket lemma av ett ord som anges och lexemnummer vilket lexem. Detta ger en tagguppsättning bestående av 13 taggar. I det webb-baserade gränssnitt som har utvecklats inom projektet för att utföra semantisk disambiguering semiautomatiskt, har texten i förväg uppmärkts med ordklass och lemma- och lexemnummer, men kan vid manuell kontroll korrigeras genom att man väljer en annan ordklass och lemma- och lexemnummer. I figur 3.8 presenteras ordet med kontext, ordklass och lemma- och lexemnummer. I figur 3.9 illustreras hur det är möjligt att välja andra alternativ än det förvalda.



Figur 3.8 Bild från disambigueringsgränssnitt.



Figur 3.9 Möjliga ordklasstaggar i första rullgardinsmenyn och möjliga lemma- och lexemtaggar i andra rullgardinsmenyn.

3.3.3 Webb-baserat gränssnitt för sökning

För att kunna göra läromedelstexterna tillgängliga i dess elektroniska skick samt i bearbetade format tillgängliga krävs att representationen av texterna och orden blir lättillgängliga. Detta görs i form av ett webb-baserat gränssnitt för sökning. En tidigare version av ett liknande gränssnitt med möjlighet till sökning i konkordansformat har utvecklats av DK. I figur 3.10 illustreras ett sökgränssnitt som används för att söka i texterna. Det finns möjlighet att ange sökord, del av ord eller helt ord. Andra tillval är läromedel, ett eller flera, val av ord i viss del av texten *texttagg*, såsom *brödtext*, *rubrik*, *bildtext* eller liknande. Vidare kan olika stilar i sökningen inkluderas eller exkluderas, t.ex. *normal*, *kursiv*, *fet* etc. I resultatet på sökningen, d.v.s. sökträffarna, som visas i en konkordans (se beskrivning i avsnittet Konkordans), kan också storlek på omgivande kontext regleras.

Figur 3.10 En illustration av sökgränssnitt för OrdiL-korpusen.

Konkordans

Att använda konkordanser för att illustrera resultat av sökträffar är numera standard i datorbaserade sökverktyg. I vårt fall används KWIC-konkordanser. KWIC står för Key Word In Context. I en konkordans, se figur 3.11, visas ett sökord i dess autentiska kontext. I exemplet anges sökordet i en mittenkolumn omgiven av både vänster- och högerkontext. Vi har också valt att ange information om vilket läromedel, ämne, texttyp och stil som utmärker ett sökord i en viss kontext.

ARTIKEL	VÄNSTERKONTEXT	SÖKORD	HÖGERKONTEXT
BIOLOGIBOK 1-Bildtext-normal	i köttstycken tillsammans med krossat	glas	. Svampar brukar delas in
BIOLOGIBOK 1-Faktatext-normal	sig en sockerbit , ett	glas	mjölk eller saft . Om
KEMIBOK 1-Bröd-normal	vatten · tungmetaller · krossat	glas	. Här är några bra
KEMIBOK 1-Bildtext-normal	är det för skillnad mellan	glas	och plast ? Vad tål
KEMIBOK 1-Bildtext-normal	är det för skillnad mellan	glas	och plast ? Vad tål

Figur 3.11 En konkordans med sökordet *glas*.

En konkordans föregås i vårt fall också av en redovisning av antal träffar i sökta läromedel. Se figur 3.12.

Korpusen KEMIBOK 1 innehåller 37 träffar
Korpusen FYSIKBOK 1 innehåller 35 träffar
Korpusen KEMIBOK 2 innehåller 20 träffar
Korpusen MATTEBOK 2 innehåller 7 träffar

Figur 3.12 Antal träffar i olika läromedel på sökordet *glas*.

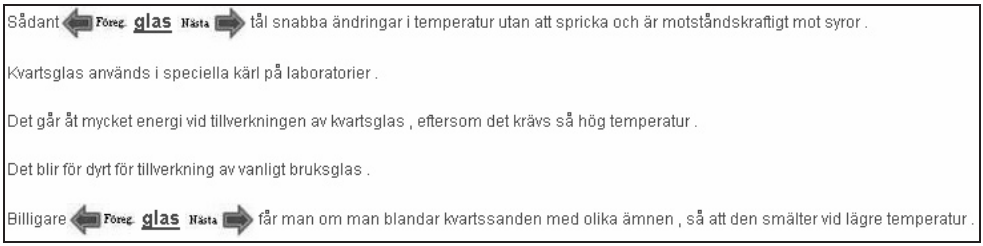
För att underlätta studiet av träffar på ett sökuttryck i konkordansformat har vi använt möjligheten att sortera informationen dels på *artikel-källhänvisning* och övrig information, *vänsterkontext*, *sökord* och *högerkontext*. Sortering utförs genom att man ”klickar” på respektive rubrik.

Full kontext (Webb-version)

Vid sökning med träffar i konkordansformat, fås endast en begränsad del av kontexten. Det kan ibland vara nödvändigt att få större kontext än då den sträcker sig över ett visst antal ord. En fullständig kontext av en sökträff fås genom att gå vidare genom den länk som finns i konkordansen om läromedel, texttyp och stil. Då skapas en Webb-version i HTML (HyperText Markup Language) i realtid av ett läromedel där en eller flera versioner av sökordet finns. Se figur 3.13 och 3.14. Då sökning sker på en del av ett ord, t.ex. ord som börjar på *de* så fås även träffar i texten som gäller *den*, *det*, *detta* etc.

I Webb-versionen av läromedlet kan alla träffar studeras eftersom de länkas ihop sekventiellt, där träff 1 länkas till träff 2 o.s.v. En träff markeras genom att den är understruken och omgiven av pilar. Genom att klicka på pilarna får användaren en möjlighet att studera nästa eller föregående träff i texten.

Figur 3.13 Uppgift om antal träffar i ett visst läromedel.



Figur 3.14 Exempel på Webb-version av ett läromedel, med markerade träffar.

3.4 Statistiska data

Ett av målen i projektet är att skapa ordlistor som beskriver ordförrådet i läromedelstexterna. Dessa ordlistor kommer att innehålla ord som är intressanta ur en statistisk synvinkel. Det kan vara ord som är flertydiga men endast används i en speciell betydelse och orden kan då också vara statistiskt snedfördelade mellan olika ämnen i läromedelstexterna. De texter som har använts som utgångspunkt för att generera statistisk information är dels alla läromedelstexter och dels en referenskorpus på ca 50 miljoner löpord bestående av G-P, DN och SDS. Vi har till största delen jämfört vårt textmaterial med text från G-P utom avsnitt från ledare, ekonomi och kultur. Se också avsnitt 2.5 i denna rapport.

I detta avsnitt ingår en del arbete som initialt utfördes av DK i form av generering av statistiska data på de fyra läromedlen som först bearbetades. För detta ändamål användes Ngram Statistics Package (NSP), version 0.57.

En annan fördel med statistiska data är att de ger en bild av förekomsten av ord och dess frekvenser fördelade på olika ämnen. Då det finns listor av ord och dess frekvens kan dessa också jämföras med varandra för att ta reda på likheter och olikheter i ordförrådet, dels inom ämnesgrupperna NO+matte och SO, men också inom enskilda ämnen.

För att ta reda på vilka ord som utmärker ett ämne eller ett läromedel så har listor med unika ord skapats. Listor består då av unika ord inom ett läromedel vilket innebär att dessa ord inte förekommer i något annat ämne eller läromedel eller i referenskorpusen. En annan metod som användes för att hitta ämnestypiska ord var att ta fram de som bestod av åtta bokstäver eller mer, se avsnitt 3.4.2. Dessutom finns det listor som innehåller gemensamma ord. De gemensamma orden finns i alla ämnen i en utvald grupp. Grupperna kan precis som listorna med unika ord bestå av NO-ämnen plus matte, SO-ämnen och dessutom en grupp som innehåller alla läromedel. Denna stora grupp jämförs då med referenskorpusen, för att bestämma vilka ord som är gemensamma för läromedelstexterna. Ord som är gemensamma för både läromedelskorpusen och referenskorpusen finns också redovisade.

Då det kan vara intressant att studera frekvenser av mer än ett ord, finns även listor med bigram – två ord som förekommer tillsammans och trigram – tre ord som förekommer tillsammans. Vid redovisning av bigram och trigram kommer ett speciellt statistiskt mått att användas, *log-likelihood*. Detta mått används för att beskriva hur sannolikt det är att två ord förekommer tillsammans i förhållande till deras inbördes frekvenser och hela textens storlek. Log-likelihood kan också användas för att jämföra ord i två olika korpusar. Se vidare avsnitt 3.4.2 i denna rapport.

I avsnitt 3.4.1 följer en beskrivning av hur statistiska data kan anges och vidare vilka listor av frekvenser som hittills har sammanställts om läromedelstexterna.

3.4.1 Hur beskrivs statistiska data?

Statistiska data kan redovisas på olika sätt. Här anges olika mått samt en kort beskrivning av i vilka situationer de är lämpliga att använda.

Man kan använda *absolut frekvens* för att ange exakt alla förekomster av ett ord.

För att kunna göra jämförelser mellan två frekvenslistor som inte innehåller exakt lika många ord är det bättre att använda *relativ frekvens*. Det är ett mått som anger absolut frekvens i förhållande till alla antal ord. Antag att relativ frekvens ska anges för ordet N och A är antal ord i hela jämförelsematerialet, då blir formeln: (Antal förekomster av N/Antal ord i hela jämförelsematerialet A) * 100= relativ frekvens i procent.

Log-likelihood är ett annat mått som anger sannolikheten för att två eller fler ord ska förekomma tillsammans i en text. Hänsyn tas till ordens inbördes frekvenser samt hela textens storlek. Detta mått kan även användas för att jämföra frekvenser av samma ord i olika korpusar.

Pearson's Chi-square (X^2) är ett mått som används för att mäta hur olika eller lika två förekomster eller fenomen är i olika texter. I beräkningen tas hänsyn till både observerad frekvens samt förväntad frekvens.

Dispersion är ett mått som används för att inte endast beskriva den observerade frekvensen utan också ange i vilken utsträckning ett ord förekommer i olika delar av en text, d.v.s. hur stor spridningen av ordet är. Men en helt jämn spridning fås värdet 1 och ju lägre värdet är desto mer snedfördelat är ordet. Måttet kan alltså användas för att sortera bort ord som inte är representativa för ett textmaterial p.g.a. snedfördelning.

3.4.2 Listor med statistiska data från läromedelstexterna (oberoende av unika och gemensamma – endast mest frekventa)

I sammanställning av frekvenslistor kommer ett antal sådana att exemplifieras. Tabell 3.5 beskriver absolut frekvens över graford inom

ett ämne. Orden representerar de mest frekventa orden inom ett ämne men är på inget sätt unika för detta ämne. De mest frekventa orden oavsett vilket ämne det gäller ser ut på ungefär samma sätt. Jämför med tabell 3.6 från ”Tiotusen i topp” som bygger på Press65-korpusen (Allén 1972).

Tabell 3.5 Alla ord inom ett ämne[9].

Absolut frekvens	Graford
2774	och
1824	att
1733	som
1675	i
1658	är
1257	kan
1179	en
1176	av
1035	på
950	till
835	med

Tabell 3.6 Tiotusen i topp.

Absolut frekvens	Graford
30856	och
29551	i
23766	att
20391	en
19610	som
18202	det
15506	är
15466	av
14668	den
14161	på
12398	för

I tabell 3.7 anges vissa graford med versaler eller gemener eller med en blandning av båda. Detta beror på att orden tolkas som textord, vilket innebär att de representeras på samma sätt som de förekommer i texten.

Tabell 3.7 Exempel på *bigram* [10] (två ord som frekvent förekommer tillsammans).

Absolut frekvens	Tvåordsenhet	
156	till	exempel
109	bland	annat
85	består	av
73	beror	på
36	orsakas	av

[9] Eftersom de mest frekventa orden inom ett ämne är vanliga generella ord, t.ex. prepositioner, så framgår det inte vilket ämne det rör sig om.

[10] De typer av uttryck som förekommer som bigram är ett bra råmaterial, men innehåller ibland en del ointressanta fall.

Tabell 3.8 Exempel på trigram (tre ord som frekvent förekommer tillsammans).

Absolut frekvens	Trigram		
46	med	hjälp	av
35	växter	och	djur
18	på	grund	av
8	djur	och	växter
7	syre	och	näring
7	ägg	och	spermier
6	stam	och	blad
6	mat	och	syre
6	bakterier	och	virus
6	armar	och	ben

3.4.3 Listor med statistiska data från läromedelstexterna (gemensamma och frekventa ord)

Tabell 3.9 innehåller ord som är gemensamma och förekommer i alla läromedelböcker i vår korpus. I tabellen anges de mest frekventa graforden i fallande frekvens.

Tabell 3.9 Exempel på gemensamma och frekventa ord.

Absolut frekvens	Graford
25132	och
20420	i
16094	att
15921	är
15231	som
13895	en
12488	av
10986	på
9896	det
8815	med

Det som kan anses vara intressant med tabeller och listor över ord som är gemensamma och förekommer i alla ämnen är att dessa ord kan vara representativa för läromedelsspråket som sådant.

3.4.4 Listor med statistiska data från läromedelstexterna (unika ord)

I tabell 3.10 anges unika ord inom ett ämne, vilket innebär att orden inte finns inom något av de andra ämnena. Exemplet i tabellen är hämtat från biologiämnet.

Tabell 3.10 Unika ord inom ett ämne.

Absolut frekvens	Graford
73	kromosomer
68	anlag
52	hjärnans
48	nerver
44	spermier
34	groddjur
34	ryggradsdjur
33	genetiska
33	könsceller
31	sporer

Att ange ord som är unika inom ett ämne gör att man också sorterar ut de ord som kan anses vara ämnesspecifika. Dessa ord kan vara till hjälp då man vill veta vilken teminologi som används inom biologiämnet.

3.4.5 Listor med statistiska data från läromedelstexterna (8-tecken eller mer)

Beroende på hur långt ett ord är kan det vara mer eller mindre vanligt i ett läromedel, jmf. också Nordman (1992) om långa ord i fackspråk. Nordman menar också att långa ord i många fall är termer. Vid en viss längd är det dessutom ofta unikt för texten/ämnet. Exemplet med ord på 8 tecken eller mer säger en del om ämnet. Se tabell 3.11. Texten är hämtad från fysikämnet.

Tabell 3.11 Utdrag från frekvenstabell med ord som är minst 8 tecken långa.

Absolut frekvens	Graford
72	strömmen
70	elektrisk
56	strålning
49	riktning
49	protoner
48	neutroner
47	omvandlas
42	stjärnor
39	hastighet
36	spänningen

Andra frekvenslistor som har skapats på materialet och är tillgängliga är listor med relativ och absolut frekvens och graford för alla ämnen, för varje ämne (två sammanslagna) och för ämnesgruppen (SO och NO) och även en lista för alla ämnen.

Relativ frekvens kan användas för att jämföra förekomster av ord eller uttryck mellan två texter oberoende av hur stora texterna är. Antag t.ex. att man vill jämföra hur ofta vissa grundämnen nämns i kemiämnet. Om man då har två olika stora läroböcker går det inte att använda absolut frekvens. Därför används relativ frekvens. Se beskrivning i avsnitt 3.4.1. I tabell 3.12 och 3.13 exemplifieras några grundämnen och deras absoluta och relativa frekvens. I tabellerna jämförs även hur frekvent användningen är av vissa passiva verb.

I tabell 3.12 och 3.13 är det möjligt tack vare uppgifter om relativ frekvens att jämföra användningen av vissa ord som beskriver grundämnen. Det framgår att *syre* och *kol* används mer frekvent i kemibok 2 medan *väte* och *kväve* är vanligare i kemibok 1. När det gäller verb i passiv form är *bildas*, *omvandlas* och *påverkas* vanligare i kemibok 2, medan *användas* och *tillverkas* är mer frekventa i kemibok 1.

Tabell 3.12 Relativ frekvens i kemibok 1

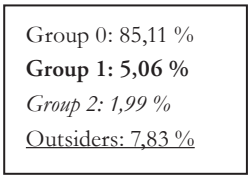
Relativ frekvens	Absolut frekvens	Graford
0.161	135	syre
0.065	54	väte
0.055	46	kväve
0.051	43	kol
0.159	133	bildas
0.041	34	omvandlas
0.053	44	användas
0.037	31	tillverkas
0.005	4	påverkas

Tabell 3.13 Relativ frekvens i kemibok 2

Relativ frekvens	Absolut frekvens	Graford
0.162	97	syre
0.045	27	väte
0.049	2	kväve
0.109	65	kol
0.273	163	bildas
0.065	39	omvandlas
0.038	23	användas
0.035	21	tillverkas
0.038	23	påverkas

3.4.6 Fixer - ett datorprogram för överblick av frekvensband i texter

Fixer är ett datorprogram, utvecklat av Martin Stissing (Århus Kommun Sprogcenter), som enkelt ger en överblick över frekvensfördelning mellan två texter. Det bygger på ett väl etablerat sätt att uttrycka frekvensband (Laufer & Nation 1995). Med hjälp av verktyget Fixer kan man jämföra en större del av ordförrådet i två texter. I exemplet nedan jämförs två texter i fysik. Man utgår ifrån hela ordförrådet i den ena läroboken, fysikbok 1, samt väljer ut ett textavsnitt från den andra boken, fysikbok 2. Jämförelse görs mellan texter på grafordsnivå vilket innebär att ord som skiljer sig kan göra det genom böjningsform eller gemener och versaler. I den sammanfattande rutan i figur 3.15 anges olika grupper där grupp 0 är ord som täcks inom de 1000 vanligaste graforden i båda texterna, dessa utgör den största delen av textmaterialen med 85,11 % av orden. Grupp 1 är de 1.001-2.000 vanligaste orden och representerar ca 5,06 % av orden. Grupp 2 är de 2.001-3.000 vanligaste orden och utgör endast 1,99 % av orden. Den sista gruppen (utomstående) är de ord som inte är gemensamma för de båda läromedlen i fysik vilket är ca 7,83 %. Sammanfattningsvis kan konstateras att det ämne inom fysik som undersöktes i fysikbok 2 handlade om *ljus*. Beskrivningen av ljus finns i båda läroböckerna. Därför är det naturligt att Grupp 0 är störst eftersom det finns många gemensamma ord i läroböckerna. Den mest intressanta gruppen är den sista, vilket innehåller ord som inte är gemensamma för fysikböckerna. Denna grupp innehåller ofta ämnestypiska och ämnesspecifika termer.



Figur 3.15 Procenttal från datorprogrammet ”Fixer” vilka anger fördelning av ord inom vissa frekvensband.

I exempel 3.1 följer dels ett utdrag ur texten, så som den markeras med färger vilka i turordning representerar grupp 0 (brödtext), grupp 1 **(fetstil)**, grupp 2 *(kursiv stil)* utomstående (understruken stil) och dels utdrag från exempel 3.2-3.5, vilka är de fyra grupper som nämns i figur 3.15

Ett öga kan *liknas* vid en kamera. Ögonlocket är ögats slutare och **pupillen** motsvarar *kamerans* **bländare**. När ljuset är starkt är **pupillen** liten och när ljuset är svagt är den stor. Efter att ljuset **passerat pupillen** träffar det ögonlinsen – *kamerans* objektiv. Ögonlinsen har till uppgift att **bryta** ljusstrålarna så att de bryts och sammanstrålar på näthinnan och skapar en bild.

Exempel 3.1 Utdrag från text analyserad av ”Fixer”.

boken, väggar, Ett, huset, modeller, verkligheten, Linserna, bilder, Annars, Objektivet, dem, håller, strålar, reflekterar, positiv, rad, bilden, kallas, synas, börjar, dåligt, uppför, snart, lätta, Allt, bestod, Konkava, hör, desto, strålarna, snabbare, ljuset, tittar, brukar, kastar, igen, annan, lite, innanför, växterna, violett, Andra, moln, blixt, ihop, fysiken, träffar, står, färre, vad, Där, infraröd, Hastigheten, vänster, sakta, marken, når, grund, samma, passerar, ett, undre, Vanligt, liknande, blandas, hålet, speciellt, Infraröd, Placera, tid, liknar, då, brännvidden, resultatet, delar, Kikare, b, ljusets, ansiktet, sorts, f, polariserat, viktigt, släpps, Man, första,

Exempel 3.2 Utdrag från grupp 0 som innehåller de 1.000 vanligaste orden.

skapas, digitala, infallsvinkel, trådar, stearinljuset, träd, Ljushastigheten, science, synfel, Solljus, därifrån, sked, nervänd, annorlunda, TV-skärm, Risken, Vattendropparna, typen, sjön, fäst, vattenyta, Resten, signal, värdet, Linjen, vitaminer, värmestrålning, stöta, ringarna, solljus, Solstrålarna, kläder, värde, riktig, delarna, tunna, sargen, bränslen, blå, vinkelrät, totalreflekteras, rep, variera, hår, pupillen, samlingslins, stearinljus, solstrålarna, blått, Tyvärr, infraröda, vägbana, skiljer, vattendroppar, bilist, stråle, vägbanan, buktar, skär, brännpunkt, vika, bländare, avbildas, tunnast, ställs, UV-strålning, r, cirka, mörka, TESTA, linje,

Exempel 3.3 Utdrag från grupp 1 som innehåller de 1.001-2.000 vanligaste orden.

CD-spelare, fickspeglar, fibrerna, blandning, beteckningen, kontaktlinser, bränglas, motsvarade, ljuslågan, grenar, läsare, gälla, möjliga, beskrivit, närsynthet, lämpliga, Inne, Fullständig, polaroidglasögon, exakt, Glasen, indelad, Hittills, molnet, 125, Färgerna, Lupp, ljuskänslig, indigo, färgade, överst, glasen, linsens, avlyssna, draget, bostaden, beräkningar, historiens, negativa, liknas, farlig, 2,5, återvänder, parabolantenn, färgen, polariserade, oförändrad, benen, inritade, försiktig, Linsens, CD, färgerna, energikälla, allvarligt, nödvändig, fotonerna, astronomisk, Okularet, bytt, fiberoptik, huvudsak, livet, lupp, Fiberoptik, glasruta,

Exempel 3.4 Utdrag från grupp 2 som innehåller de 2.001-3.000 vanligaste orden.

Blundade, CD-skiva, nedsänkningar, Infalls, skapa, brytningsvinkel, förbränning, spegelvänts, stämningsfull, 1500, gastroskop, grekiskans, brännpunkter, bearbetas, ljuslågans, Ozonskiktet, idag, Idag, dioptrier, spetsen, soligt, hejdas, halvt, svängs, frigöra, köpa, utföranden, målarfärgen, avskärmningsanordning, fungeran, lurat, bländaren, energirikt, dåliga, ringa, ursprungligen, datorer, skydd, luras, fiberoptiken, blåa, t.ex., Slutarens, färdats, avhjälpa, märkliga, skenbild, fiber, strålningstyperna, hastighetsmätare, lurade, plåta, staket, ljusstrålen, verklig, stärkelse, konvergenta, Vattenvågor, ljusförhållandena, mulet, äter, Färgbanden,

Exempel 3.5 Utdrag från utomstående ord, d.v.s. de ord som skiljer sig mellan de två fysikämnena.

Referenser

Allén, S. 1970, 1971. *Nusvensk frekvensordbok (NFO), Del 1-3*, Data linguistica, Almqvist & Wiksell, Stockholm.

Allén, S. 1972. *Tiotusen i topp*, Data linguistica, Almqvist & Wiksell, Stockholm

Berg, S. 1970. *Homografi i nusvenskan. 1. Studier över homografi. 2. Svenskt homografiexikon*. Institutionen för nordiska språk vid Göteborgs universitet. Göteborg maj 1970.

Brants, T. 2000. TnT - *A Statistical Part-of-Speech Tagger*. I *Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000*, Seattle, WA.

Daelemans, W. 1999. *Machine Learning Approaches*, i van Halteren (ed.), *Syntactic Wordclass Tagging*, Kluwer, Academic Publishers.

Garside, et al. 1997. *Corpus Annotation, Linguistic Information from Computer Text Corpora*, Addison Wesley Longman Inc., New York, s.21-24,

Johansson Kokkinakis, S. 2003. *En studie över påverkande faktorer i ordklasstaggning. Baserad på taggning av svensk text med EPOS*. Avhandling, Humanistiska fakulteten, Göteborgs universitet.

Järborg, J. 2003. *Formaliserade semantiska samband mellan enheter i GLDB*. Research Reports from the Department of Swedish, Göteborg University. GU-ISS-03-1.

Kokkinakis, D. 2001. *A Framework for the Acquisition of Lexical Knowledge; Description and Applications*, s.69-72, Avhandling, Humanistiska fakulteten, Göteborgs universitet.

Kokkinakis, D. 2004. *Reducing the Effect of Name Explosion*. Proceedings of the LREC Workshop: Beyond Named Entity Recognition, Semantic labelling for NLP tasks. Fourth Language Resources and Evaluation Conference (LREC). Lissabon, Portugal.

Källgren, G. 1998. *Documentation of the Stockholm-Umeå Corpus* (DRAFT), Stockholms universitet.

Laufer, B. & Nation, P. 1995. *Vocabulary Size and Use: Lexical Richness in L2 Written Production*, Oxford University Press.

Nordman, M. 1992. *Svenskt fackspråk*, Studentlitteratur, Lund.

Svenska Akademien, 1999. *Svenska akademiens grammatik (SAG)*, MediaPrint, Uddevalla.

Datorprogram och webb-länkar

Brants TnT-tagger, 2005, www.coli.uni-saarland.de/~thorsten/tnt. (060606).

Ngram Statistics Package (NSP), version 0.57, Ted Pedersen, Department of Computer Science, University of Minnesota, Duluth

SPSS for Windows, Rel. 11.0.1. 2001. Chicago: SPSS Inc.

XML, www.w3.org (070219)

4. Långa ord – en svårighet för flerspråkiga studerande?

Margareta Holmegaard

4.1 Inledning

Den undersökning som här presenteras syftar till att utpröva olika modeller för diagnostisk bedömning av ordförråd med utgångspunkt från OrdiL-korpusen. Främst avses att kunna skatta ordförrådet hos flerspråkiga, d.v.s. elever som anger att deras modersmål inte är svenska utan annat/andra modersmål. Sådana diagnostiska test kan användas för att identifiera en elevs skolrelaterade ordförråd såväl på ett mer generellt plan som i förhållande till ett visst ämne. Testresultaten kan utnyttjas för planering av en mer systematisk språkutvecklande undervisning och också bidra till att såväl lärare som elever ökar sin medvetenhet om svårigheter förknippade med ordförrådet i ämnesundervisningens läromedel. I följande forskningsbakgrund presenteras undersökningar som har fungerat som inspiration och incitament till denna studie.

4.2 Forskningsbakgrund

Det är väl dokumenterat genom olika studier att storleken på en läsares ordförråd har stor inverkan på hur väl man kan tillgodogöra sig innehållet i en text (se t.ex. Read 2000). En av flera faktorer som bidrar till att en text blir svår att läsa och förstå är andelen ord som innehåller många stavelser. I studier av engelska texter har man ofta använt en enkel formel baserad på två variabler, nämligen ordfrekvens och meningslängd (ibid. s.191). Enligt en hypotes som bygger på att långa ord är relativt ovanliga i texter och därmed svåra för läsaren skulle ett mått på ordlängd kunna ge en indikation på en texts läsbarhet. När andraspråkslärare ibland har använt sig av tester som mäter ordfrekvens och meningslängd för att avgöra en texts svårighetsgrad, har detta emellertid kritiserats av bl.a. Carrel (1987) som menar att många andra faktorer kan påverka en texts tillgänglighet för andraspråksläsare som t.ex. kulturell kunskap, textuppbyggnad och genre. Även modersmål anges av Brown (1997:90) som en viktig faktor. Om inläraren har ett modersmål, som är nära släkt med målspråket, har dem en fördel jämfört med andra inlärare, eftersom många ord är liknande.

I Sverige är den mest kända läsbarhetsformeln C.H. Björnsons LIX (Björnson 1968, se även Melander 2004:33f.). Detta läsbarhetsindex beräknas genom att man adderar den procentuella andelen ord som är längre än sex bokstäver, och den genomsnittliga meningslängden mätt i antalet ord per mening. LIX har använts flitigt i undersökningar av lärobokstexter. Melander menar dock att man inte kan nå någon djupare förståelse av läroböckernas språk enbart genom detta värde utan att även textuella egenskaper behöver studeras för att ge en mer fullständig bild. LIX-värdet kan emellertid ge en viss insikt om huruvida en text är avancerad och behöver kompletteras med andra undersökningar.

För att mäta avkodningsförmåga och semantisk kompetens av betydelse för läsfärdighet har flera tekniker använts av olika läsforskare (Lundberg 1984, Miller-Guron 1999, Adams 1990, Connor & Olsson 1990). Enligt Connor & Olsson är en god och säker avkodning en grundförutsättning för läsförståelse men långt ifrån någon garanti eftersom god språkförståelse också är en nödvändig

förutsättning. För att identifiera elever med lässvårigheter av olika slag har man t.ex. använt tester för att få kunskaper om hur inlärare kan dela ord i betydelseenheter (Johansson 1999). Johansson redogör för longitudinella undersökningar med en form av ordkedjetest och meningskedjor, där man delar ord och hopskrivna meningar i stavelseenheter. Dessa har givit signifikanta resultat vad gäller avläsningsteknik och avläsningsproblem och verkar kunna användas för att identifiera lässvaga elever. Under åren 1997 och 1998 gjordes en undersökning på totalt 400 elever i år 6-9 (Jacobsen 1993). Eleverna fick här under tidspress dela in ord i meningsbärande enheter. Syftet var att undersöka språklig medvetenhet och i vilken utsträckning eleverna kunde identifiera ordgränser enligt den ortografiska strategin. Vissa feltyper visade sig vara relaterade till avkodningsproblem medan andra hade att göra med mer allmän språklig kompetens. Ett annat resultat som gällde elevernas egen insikt om sin läsförmåga visade att flera elever med uppenbara lässvårigheter ansåg att testet var lätt och alltså överskattade sin förmåga (Johansson 1999:61). Slutsatsen från denna undersökning med ordkedjor och meningskedjor var att man med hjälp av sådana test relativt snabbt och lättadministrerat kunde identifiera elever med lässvårigheter. Det påpekas dock att orsaken till svårigheter inte kunde klarläggas enbart med dessa tester. En djupare undersökning av kunskaperna hos de elever som erhöll dåliga testresultat rekommenderades därför.

I flera skolor har ord- och meningskedjetest också använts för att identifiera elever med dyslexi. Flerspråkiga elever, som inte har hunnit skaffa sig kunskaper i svenska, har många gånger tillhört de elever som gjort svaga testresultat på ordavkodning i dessa test (Johansson 1999:65f). Eftersom man sällan närmare undersökt orsakerna till elevernas svårigheter har de inte alltid fått adekvat andraspråksundervisning och/eller bedömts vara i behov av specialundervisning. Alltför många flerspråkiga elever bedöms således på osäkra grunder och effektivare testmetoder behövs (Skolverket 2005, Taube 2000). Det finns ett stort behov av olika test för att diagnostisera flerspråkiga elevers läsförmåga och ordförståelse, inte minst för att kunna erbjuda dem behovsanpassad och effektiv undervisning.

4.3 Syfte med undersökningen

Med hjälp av information från OrdiL-databasen har fyra testmodeller konstruerats för forskningsändamål. Dessa är tänkta att även kunna komma till praktisk användning för lärare och elever i undervisningen. Undersökningen syftar till att utforma modeller för diagnostiska prov som dels kan identifiera elevers allmänna vokabulär ur såväl kvantitativa som kvalitativa aspekter, dels kan ge kunskap om vilka svårigheter elever kan ha i samband med ämnesundervisningen då de ska tillägna sig kunskap via texter och läroböcker. Enligt bl.a. PISA-undersökningarna (2001, 2003) vet vi att läsförmåga och förståelse av ord även påverkar läsförståelsen i fackämnen såsom matematik och naturorienterande ämnen. Detta gäller troligtvis även samhällsorienterande ämnen där läromedlen spelar en stor roll i undervisningen.

De test som här utprövats fokuserar på ord i samhällsorienterande ämnen (SO-ämnen), men resultaten av studien kan förhoppningsvis vara överförbara även till andra ämnen. I det följande redogörs för ett flerdelat testförsök med långa ord i lärobokstexter där ordkunskapen testas genom 1. *identifikation av betydelsedelar* (Test A), 2. *självskattning* (Test B), 3. *matchningstest* (Test C) och 4. *flervalstest* (Test D). Testen fokuserar framförallt receptiva aspekter av ordkunskapen (Test A, C och D), men i självskattningstestet (Test B), aktualiseras även produktiva aspekter, eftersom informanterna kan ha olika uppfattningar om vad det innebär att ”kunna ett ord”. Det bör framhållas att ordförrådet inte strikt kan delas in i dessa två kategorier. Snarare kan man se på ordförrådet som ett receptivt-produktivt kontinuum, som återspeglar olika grader av kunskap och förtrogenhet med orden i ett språk (Melka 1997).

Undersökningen genomfördes under vårterminen 2005 i tre olika elevgrupper, dels i en grundskola och en gymnasieskola med hög andel tvåspråkiga elever dels på universitetet i en studerandegrupp med utländsk gymnasiekompetens.

Syftet med testen var att undersöka följande frågor:

- Kan informanterna dela de utvalda SO-orden i komponenter efter deras huvudsammansättningsled (Test A)?
- Hur skattar eleverna sin egen kunskap om orden (Test B?)
- Kan informanterna koppla ett antal testord till rätt förklaring (Test C och D)?
- Finns det samband mellan resultaten av de olika testerna?
- Finns det skillnader i resultaten mellan elever med svenska som modersmål och övriga elever?
- Finns det skillnader mellan könen?
- Finns det skillnader mellan elever i olika skolformer, med olika skolbakgrund?
- Kan testen användas för att identifiera elever med olika språkfärdigheter?

4.4 Metod

4.4.1 Uppläggning av testen

I följande undersökning har fyra olika test använts för att mäta olika aspekter av lexikal kompetens hos informanterna. Samtliga test utgår från 57 ord som innehåller flera stavelser. De fyra testen har utformats efter fyra olika principer:

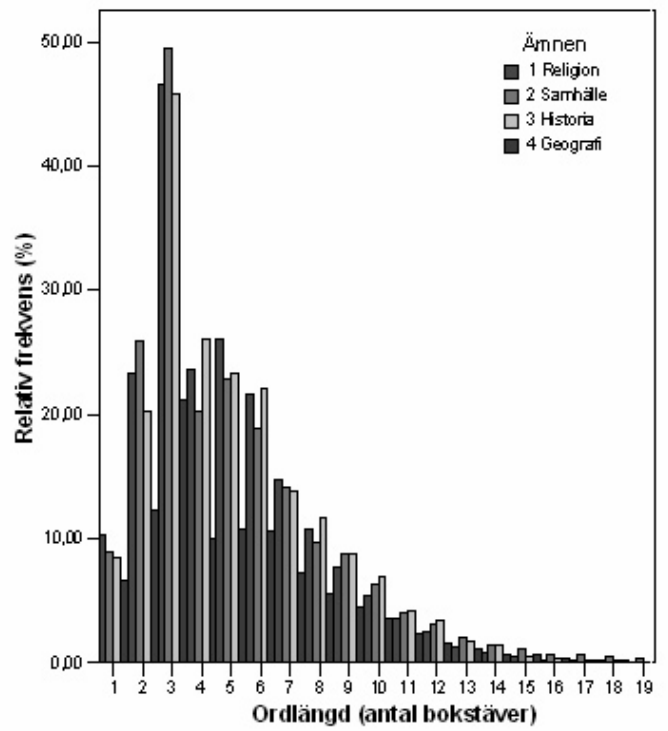
- Indelning av de 57 testorden i betydelsedelar (Test A). Här ska orden delas i två delar efter deras huvudsammansättning (Bilaga 2).
- Självskattningstest av de 57 testorden (Test B). Här ska informanterna göra en självskattning av sin förståelse av de ord som de tidigare mött i test A utifrån tre alternativ: vet säkert, vet inte alls, är osäker (Bilaga 3).
- Matchningstest (Test C). Här har nio av de 57 testorden använts i en matchningstest där informanterna ska koppla ihop testorden med alternativa förklaringar av orden (Bilaga 4).
- Flervalstest. Här har sex av de 57 testorden använts i ett flervalstest där informanterna ska para ihop dem med korrekt förklaring bland fyra alternativa svar (Bilaga 5).

Det första testet (Test A), där ord ska indelas i betydelsedelar bygger på antagandet att det är lättare att förstå ett långt ord om man kan dela in det i dess huvudbeståndsdelar eftersom betydelsen av ordens olika delar då kan antas vara bekant. Om inläraren har svårigheter att se hur orden är uppbyggda borde det utgöra en extra belastning vid läsningen av en text. Självskattningstest (Test B) anses vara ett bra komplement vid språktestning (Oscarson 1999) vars resultat korrelerar med andra typer av test som mäter språkkunskaper. Självskattning anses också förhöja intresset för ämnet och ha en positiv effekt på själva inlärningen (von Elek 1981, 1985). Enligt Shrauger & Osberg (1981) har relativt avancerade inlärare många gånger en bra uppfattning om sina språkfärdigheter och kan bedöma sina kunskaper bättre än en utomstående. Självskattning påverkas emellertid av olika faktorer exempelvis skolbakgrund, förväntningar, ambitionsnivå och vana vid självstudier och egen bedömning. Även kulturella skillnader och återkoppling på testen (Bachman & Palmer 1989, Lewkowicz & Moon 1985, Strong-Krause 1997) kan påverka resultaten. Paul Meara som har arbetat med test enligt självskattningsmetoden (Meara & Jones 1988, Meara 1996) menar att det är ett effektivt sätt att snabbt mäta kunskap om ett större antal ord. Självskattning mäter visserligen endast vissa sidor av ett receptivt ordförråd, men ger ändå en uppfattning av hur många ord en person känner igen.

Syftet med Test C och D är att undersöka den receptiva ordkunskapen och igenkänningen av ord i en kontext för att på så sätt få närmare information om hur informanterna förstår orden. Vid konstruktionen av dessa test har stor vikt lagts vid förklaringarna av de utvalda orden som måste vara tydligt och enkelt formulerade, så att informanterna inte svarar fel på grund av att de har svårt att förstå förklaringarna. Val av relevanta förklaringar har gjorts med utgångspunkt från konkordanser ur OrdiL-korpusen och även med hjälp av Svensk skolordlista (2004).

4.4.2 Urvalet av testord

I OrdiL-korpusen är långa ord relativt sett ovanliga (se figur 4.1). Det gör dem också många gånger svåra att förstå (Harrison (1980:79f). Det är därför intressant att undersöka förståelsen av långa ord i läromedel för att se i vilken utsträckning de utgör problem för elever med olika språklig bakgrund. Uppbyggnaden av långa ord är i svenskan till skillnad från i t.ex. engelskan ofta genomskinlig, d.v.s. grundorden går tydligt att urskilja, vilket underlättar förståelsen av orden förutsatt att man känner till principen för ordbildning i svenskan.



Figur 4.1 Långa ord i SO-ämnen i OrdiL-korpusen.

Som underlag för denna studie har de längsta (ej lemmatiserade) orden i de olika läromedlen i historia, samhällskunskap, religion och geografi för skolår 8 sammanställts ur OrdiL-korpusen. Dessa ord består av 14-24 bokstäver och innehåller 4-8 stavelser. Av dessa val-

des 57 testord ut varvid de mest typiska fackorden undveks. Orden kan ändå betraktas som typiska för ämnena och kan lätt identifieras ämnesvis t. ex *högmässogudstjänst, frälsningssoldat* (religion) och *regeringskris, arbetsmarknadsutskott* (samhällskunskap). Avsikten med att inte ta med alltför typiska fackord var att eleverna skulle ha en möjlighet att känna igen någon del i de långa orden utan speciella fackkunskaper. För att testet skulle bli mindre omfattande koncentrerades detta test endast till SO-orden. Avsikten var att det skulle kunna utföras i klassrummet under en lektion och då endast uppta cirka 30–40 minuter av undervisningstiden.

4.4.3 Undersökningsgrupperna

Tre elevgrupper deltog i undersökningen som omfattade totalt 87 elever, 55 flickor/kvinnor och 32 pojkar/män. Urvalet gjordes med tanke på att utprövningen av testet skulle genomföras i grupper med olika bakgrundsvariabler. Grupp 1 utgörs av 43 elever i år 8 vid en invandrartät grundskola i Göteborg. Grupp 2 består av 31 elever från två omvårdnadsklasser i en invandrartät gymnasieskola. I grupperna ingår både enspråkiga och flerspråkiga elever. Grupp 3 består av 13 studerande med utländsk bakgrund som läser en intensivkurs i svenska som andraspråk med siktet inställt på att läsa vidare på universitet. De deltar i en speciell utbildning där de förutom svenska som andraspråk (sva) läser engelska och matematik och fått tillgodoräkna sig övriga ämnen från tidigare skolgång. De har alla gymnasiekompetens och sin huvudsakliga skolgång i hemlandet. De omnämns här i undersökningen som universitetsstuderande[11].

4.4.4 Bakgrundsenkät

För att få kunskap om vissa bakgrundsuppgifter om eleverna utformades en enkät (Bilaga 1a) med utgångspunkt från följande variabler:

[11] Vissa av informanterna genomförde ochså denna kurs på universitetet.

språk, födelseår, kön, modersmål, antal år i Sverige, skolgång i Sverige, skolgång i hemlandet, antal skolår, deltagande i svenska som andraspråksundervisning, svenskundervisning och/eller modersmålsundervisning

4.4.5 Genomförande

Eftersom det var av intresse både för lärarna och eleverna att få genomföra testen i samband med undervisningen utgjorde dessa en del av elevernas reguljära arbete inom SO-ämnena och genomfördes av undervisande lärare. Att lära ord och testa ordkunskap i de olika fackämnena i samband med ämnesundervisning är säkerligen mer effektivt än i en renodlad språkundervisning och genom allmänna språktest (Read 2001:159).

De studerande fyllde först i bakgrundsenkäten och fick därefter under tio minuter dela in testorden (Test A) i deras två huvuddelar som i exemplet: serie/tidning, tidnings/läsning. Avsikten med tidsbegränsningen var att de studerande skulle arbeta relativt snabbt, men inte så snabbt att de upplevde situationen stressande. I test B, C och D fick informanterna arbeta i egen takt och hela testförfarandet genomfördes under cirka 30-40 minuter. Testsvaren har därefter lagts in och databearbetats med hjälp av statistikprogrammet SPSS.

4.5 Resultatredovisning

4.5.1 Bakgrundsenkäten

I bilaga 1b finns bakgrundsfakta för gruppen hämtade från enkätfrågor (Bilaga 1a) med avseende på modersmål, deltagande i olika skolformer och kön. I gruppen som helhet är 55 av de 87 informanterna kvinnor. Informanterna är födda mellan 1984 och 1990. Om man studerar uppdelningen på de olika skolformerna ser man att av eleverna på grundskolan är 38 födda 1989 medan 5 är födda 1990. På gymnasiet är 11 studerande födda år 1984-1987, 17 studerande 1988 och 3 studerande år 1989. De som är födda mellan 1984 och 1987 går år 1 i gymnasieskolan

trots sin ålder. På universitetet är informanterna födda 1984-1987 och de har alla gymnasiebehörighet från sina respektive hemländer.

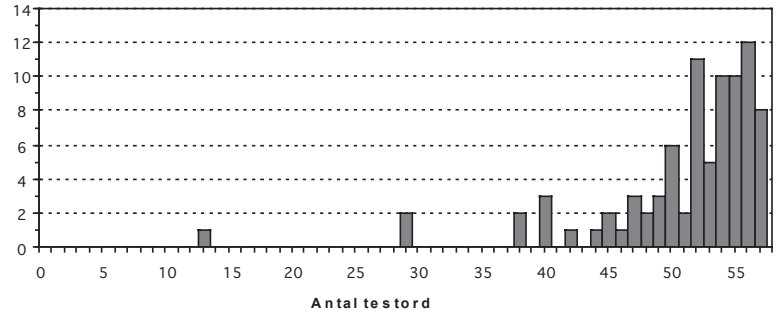
Flertalet elever på grundskolan är födda i Sverige och har gått hela grundskolan i svensk skola. De elever som kommit under grundskoletiden har mycket olika utbildningsbakgrund. En elev i grundskolan anger t.ex. att hon endast gått fem år i skolan totalt. I gymnasieskolan har fem elever gått kortare tid än brukligt i skolan medan åtta elever har angett längre skoltid. Dessa elever har kommit till svensk skola under högstadietiden och har då behövt gå några extra år för att lära sig svenska. De studerande vid universitetet har alla en gedigen skolgång i hemlandet och har endast gått i svenskt skolsystem 1-2,5 år.

Det är ofta svårt att få korrekta uppgifter om elevers deltagande i undervisning i svenska, svenska som andraspråk och modersmål. Det har bl.a. att göra med att det i vissa skolor inte görs någon skillnad mellan ämnena svenska och svenska som andraspråk. På grundskolan deltar 11 av de 18 elever som har annat modersmål än svenska i sva-undervisning medan 6 av eleverna med annat modersmål än svenska uppger att de deltar både i sva- och i svenskundervisning. 6 elever deltar i modersmålsundervisning. I gymnasieskolan deltar flertalet, 21 av de 24 elever som angett annat modersmål än svenska, i sva-undervisning, medan en elev deltar i både sva- och svenskundervisning. 8 av dem deltar även i modersmålsundervisning. Man kan således notera att flera elever som angett annat modersmål än svenska på grundskolan deltar i vanlig svenskundervisning medan de flesta på gymnasiet deltar i andraspråksundervisning. På båda stadierna deltar cirka en tredjedel i grundskolan och gymnasieskolan i modersmålsundervisning. På universitet läser samtliga svenska som andraspråk.

I gruppen som helhet finns totalt 22 språk representerade. Av eleverna har 25 angett svenska som sitt modersmål, 7 har inte besvarat frågan, 3 informanter har uppgivit mer än ett modersmål, de övriga har varit fördelade på 21 olika språk. De sju elever som avstår från att ange något modersmål kan ha känt osäkerhet över vilket språk de skulle välja eftersom de inte visste om de kunde ange flera språk som modersmål.

4.5.2 Test A Indelning av testord i betydelsedelar

Resultatet av Test A där informanterna ska dela orden i två delar efter deras huvudsammansättningar (enligt figur 4.2) visar att den informant som har minst antal rätt på test A har 13 poäng av 57 möjliga. Nästa informant har 29 poäng. Det innebär att av de 85 deltagare som gjort test A (2 har ej fullföljt) har alla utom en klarat minst hälften av uppgifterna. Medelvärdet är 51, vilket är högt på en skala med 57 poäng. Testet har en ”tak-effekt”, där de välpresterande inte kan visa sin förmåga fullt ut. Det innebär att testet kunde ha gjorts svårare eller givits kortare tid. Enligt liknande test med ordkedjor (Johansson 1999) är tidsfaktorn avgörande när man ska mäta läsfärdigheten för att få fram signifikanta resultat.



Figur 4.2 Antal poäng som markerar rätt i test A.

Även om informanterna behärskar delning av orden relativt väl förekommer ett antal avvikelser. De felaktigt indelade orden i test A, d.v.s. inte enligt huvudordsprincipen, kan fördelas i olika kategorier. I tabell 4.1 listas de ord som felmarkerats hos fler än en informant. De anges i procent i fallande frekvens.

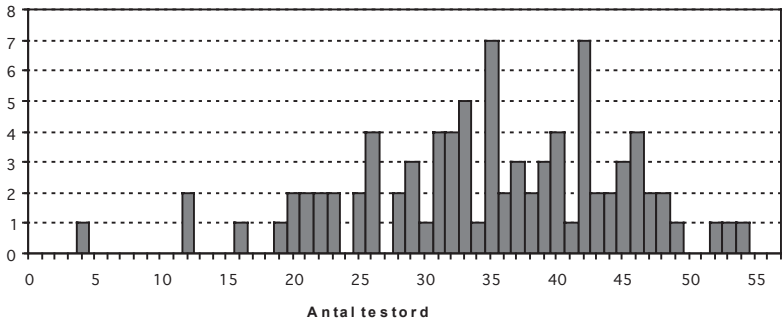
Tabell 4.1 Felmarkerade testord (test A).

Testord	Felmarkering i procent
motreformation	63
sammanslagning	33
justitiedepartement, massarbetslöshet, valsensation	21-28
gästgivaregård, konfirmationsakt, riksdagsledamöterna, samhällsinflytande	15-16
arbetsmarknadsutskottet, oppositionspartier, majoritetsbeslut	12-14
nattvardsfirande, reformationsrörelse	11
kommunfullmäktige, järnvägsknutar, partitillhörighet, produktionsmedlen, utbildningsutskott	10

Det i särklass svåraste ordet visar sig vara *motreformation*. Många informanter har här inte kunnat urskilja de två huvudkomponenterna utan analyserat ordet på ett sätt som tyder på igenkänning av ordet *formation*, d.v.s. ”motre-formation”, vilket inte kan antas vara till någon ledning när det gäller att förstå det här aktuella ordet. En annan typ av felaktig analys har att göra med svenskans s.k. foge-s. Man kan förmoda att många känner till regeln att många sammansatta flerstaviga ord hålls ihop med ett foge-s och därför rutinmässigt analyserat ordets huvuddelar i enlighet med denna regel, vilket kan leda till helt felaktiga tolkningar som: ”vals-ensation” och ”sammans- lagning” i stället för *val-sensation* och *samman-slagning.* Andra ord som analyseras på ett avvikande sätt kan tyda på att informanterna inte känner igen ordets huvuddelar eller identifierar andra i sammanhanget irrelevanta enheter: ”kommunfull-mäktige”, ”gäst-givaregård”, ”kon-firmationsakt”.

4.5.3 Test B Självskattning

Resultatet av självskattningen i Test B avviker från resultatet i Test A genom ett avsevärt lägre medelvärde och en större spridning i resultatet. Här markeras de ord som informanterna säger att de ”säkert vet” enligt följande:



Figur 4.3 Antal ord som informanterna säger sig ”säkert veta” betydelsen av.

Resultatet av Test B (figur 4.3) uppvisar en spridning från 4 till 55 poäng. Variationen är således stor. Medelvärdet är 35. Vid en beräkning i test B av Cronbachs alfa visar värdet 0,977 och reliabiliteten är således mycket god[11].

Pojkarna skattar sig högre än flickorna men överlag tycks eleverna göra en relativt realistisk bedömning av vad de kan (jämför avsnitt 4.6.2).

I tabell 4.3 listas ord markerade med svarsalternativet ”vet inte alls”. Det är påfallande att flera av de ord som enligt självskattningen upplevs helt okända, tillhör religionsämnet. Detta är ord som elever som är uppfödda i Sverige kan stöta på i sin vardag men som för elever med annan kulturell bakgrund och med andra religioner än den kristna kan vara främmande exempelvis: *begravningsceremoni, församlingsbor, konfirmationsakt, gudstjänstbesökare, nattvardsfirande, högmässogudstjänst.* Andra ord som kanske enbart påträffas i religionsämnet: *munkkloster, vigselförrättare, reformationsrörelse, samfundsgränser* kan vara främmande även för ungdomar med svensk bakgrund i dagens sekulariserade samhälle där få ungdomar konfirmeras och besöker gudstjänster.

[11] Cronbachs alfa är ett mått på den typ av reliabilitet som kallas intern konsistens. Graden av intern konsistens avser hur väl de olika delarna av testet mäter samma egenskap, d.v.s. i detta fall hur de olika delorden i testen var och en för sig alla mäter elevernas ordkunskap på ett liknande sätt. Reliabilitetsvärdet bör alltid vara över 0,7 för att testet ska anses ha en acceptabel reliabilitet. Ju närmare värdet 1,0 desto bättre är sambandet mellan de olika frågorna i testet.

Tabell 4.2 Ord som markerats med svarsalternativet ”vet inte alls vad ordet betyder”.

Testord	Andel informanter som uppgivit att de inte känt till ordet, i %
Begravningsceremoni	8
Sammanslagning, motreformation, munkkloster	7
Ställföreträdare, valsensation, församlingsbor	6
Ungdomsverksamhet, nattvardsfirande, Reformationsrörelse, konfirmationsakt	5
Högmässogudstjänst, vigselförrättare, samfundsgränser	3
Musikgudstjänst, gudstjänstbesökare, utrikespolitik, Parlamentsledamöter, utbildningsutskottet, yttrandefrihetsgrundlagen,åldersfördelning, tryckfrihetsförordningen, arbetsmarknadsutskottet	2 1

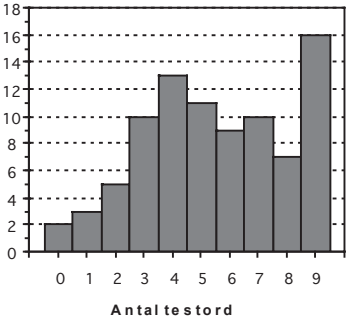
Andra ord som refererar till samhällsföreteelser är okända för flera informanter: *ungdomsverksamhet, utrikespolitik, parlamentsledamöter, tryckfrihetsförordning* m.fl. Det kan vara värt att notera att t ex ordet *ungdomsverksamhet,* som man kan förmoda att många ungdomar stött på, ändå upplevs svårförståligt. Ordet *verksamhet* tillhör den grupp av abstrakta ord med generell betydelse som kanske inte alltid uppmärksammas i undervisningen och som därför kan vålla problem för andraspråkselever.

Andra sammansatta ord kan på grund av att de innehåller konsonanten ”s” vara svåra att analysera. Här förekommer feltolkningar genom att eleverna missleds att tolka ”s-et” som foge-s i stället för en slut- eller initialkonsonant i de ingående orden. Detta kan alltså vara förklaringen till att orden *sammanslagning* och *valsensation* markerats som okända.

4.5.4 Test C Matchningstest

Resultatet av Test C där informanterna ska koppla ihop testorden med alternativa förklaringar av orden visar att antalet godkända/korrekta svar

varierar mellan 0 och 9. Medelvärdet är 5,5. Reliabiliteten avseende intern konsistens i testet är god med ett värde på Cronbachs alfa = 0,85.



Figur 4.4 Antal korrekta svar i test C

Flertalet ord i test C kan nog betraktas som relativt vanliga i samhället och informanterna kan träffa på dem i såväl i samhällsdebatten som i sin egen vardag. Trots detta är många informanter obekanta med flera ord såsom *oppositionspartier, trafikföreskrifter, fackföreningsrörelse* och identifierar inte rätt svarsalternativ i de olika meningarna som presenterats som förklaring (tabell 4.3).

Tabell 4.3 Ord som markerats korrekt (anges i procent).

Testord	**Andel korrekta svar i %**
Gudstjänstbesökare	88
Utbildningsutskott	76
Riksdagsledamöterna	65
Majoritetsbeslut	59
Förhandlingsvilja	55
Ställföreträdare	52
Oppositionspartier	49
Trafikföreskrifter, fackföreningsrörelse	48

Man kan notera att orden *ställföreträdare, gudstjänstbesökare* och *utbildningsutskott* av många elever betraktades som okända i självskattningen men då de i detta test presenteras i kontext inte tycks vålla lika stora problem. Korrelationen mellan resultaten i självskattningen i Test B och matchningstestet Test C är låg. (Se vidare avsnitt 4.6).

De nio testordens förekomst i läromedlen framgår av tabell 4.4 där också de ingående huvudledens förekomst redovisas:

Tabell 4.4 Frekvenser för testorden i OrdiL-korpusen.

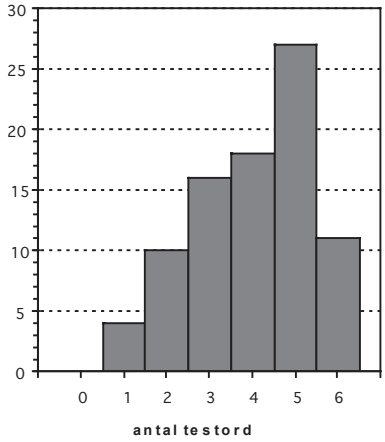
Testord	Antal förekomster		
	Hela ordet	Första huvudledet*	Andra huvudledet*
Gudstjänst/besökare	3	76	10
Utbildnings/utskott	2	48	16 (8)
Riksdags/ledamötern	1	440	93
Majoritets/beslut	1	23	93
Förhandlings/vilja	1	20	53
Ställ/företrädare	1	5 (333)	2
Trafik/föreskrifter	2	51	6
Fackförenings/rörelse	1	11	20 (125)
Oppositionspartier	1	**3**	103

* polysemt ord anges inom parentes

De fyra svåraste orden för informanterna (se föregående tabell och här de 4 sista testorden) visar sig också vara sällsynta i lärobokskorpusen, vilket framgår av tabell 4.4. Man kan notera att de som sammansättningar används endast vid enstaka tillfällen i läromedelskorpusen och att även huvudleden sällan förekommer som separata ord (förutom *trafik* och *partier*). Särskilt intressant är huvudleden *ställ-* och *rörelse*, som är polysema och mycket vanligare i andra betydelser än den som avses i testordet. Ordformen r*örelse* är ofta relaterad till verbet *röra sig* och ordformen *ställ* till imperativen *ställ* eller andra avledningar av verbet *ställa*. Detta verkar utgöra en svårighet när man ska tolka orden.

4.5.5 Test D Fl**er**valstest

Som framgår av figur 4.5 är deltagarna fördelade över hela skalan men en majoritet av eleverna har en relativt hög andel godkända svar. Medelvärdet ligger på 4. Reliabiliteten avseende intern konsistens visar ett Cronbachs alfa på 0,44. Det låga värdet beror förmodligen på att alltför få testord ingått i testet.



Figur 4.5 Antal korrekta svar i test D.

Flera elever klarar av att matcha de utvalda orden med hjälp av kontexten. Testordens svårighetsgrad framgår av tabell 4.5.

Tabell 4.5 Andel korrekta svar efter testord.

Testord	Andel korrekta svar i %
Vigselförrättare	84
Massarbetslösheten	75
Kommunfullmäktige	74
Tillverkningsindustri	64
Justitiedepartement	62
Samhällsinflytande	38

Ordet *vigselförrättare* kände flera till med hjälp av kontexten. *Massarbetslösheten* var ett ord som flertalet informanter kunde förstå. De är för övrigt det enda ord där resultaten i samtliga test stämde överens. Värt att notera är att ordet s*amhällsinflytande och tillverkningsindustri* fanns bland de ord många informanter inte kände till i test D. Svårigheten med dessa ord för andraspråkselever kan kanske tillskrivas det faktum att grundbetydelsen av *inflytande* och *tillverka* är okänd, vilket visar att även vanliga och allmänna ord kan vara oklara för dessa

elever. Förmodligen förklaras inte heller denna typ av ord så ofta i ämnesundervisningen. Ord som j*ustitiedepartement* och *kommunfullmäktige* är däremot svåra då de snarast är att betrakta som fackord och kräver goda kunskaper om samhället. Informanter med kunskaper i andra västerländska språk kanske har lättare att förstå dessa ord, eftersom de kan känna igen delar av orden men för övriga elever kan orden säkert upplevas som främmande.

En frekvenssökning av testorden i OrdiL-korpusen gav följande resultat:

Tabell 4.6 Frekvenser för testorden i OrdiL-korpusen.

Testord	Antal förekomster		
	Hela ordet	**Första huvudordet***	**Andra huvudordet**
Samhälls/inflytande	1	233	28
Justitie/departement	1	8	37
Tillverknings/industri	1	83	54
Kommun/fullmäktig	78	53	9
Mass/arbetslöshet	1	134 (76)	23
Vigsel/förrättare	1	6	0

*polysemt ord anges inom parentes

Ju oftare man möter ett ord desto lättare torde det vara att känna igen och förstå ordet. Man kan notera i tabell 4.6 att samtliga ord utom *kommunfullmäktige* förekommer ytterst sparsamt i korpusen. Om man däremot delar orden i huvudsammansättningar blir det något annorlunda. *Inflytande* och *justitie-* som uppträder relativt sällan är också svåra för informanterna. *Arbetslöshet* som inte förekommer särskilt ofta som enskilt ord torde ändå vara lätt att förstå, då man kan förmoda att det är ett ord som man hör i samhällsdebatten. *Vigselförrättare* förekommer sällan, men har ändå förmodligen med hjälp av kontexten varit lätt att tolka. I självskattningen anger däremot flertalet informanter att de inte förstår ordet.

4.6 Korrelation mellan testen

Tabell 4.7 visar hur resultaten på de olika delproven korrelerar med varandra. Detta kan indikera hur lika och olika delproven är avseende att mäta olika förmågor och färdigheter på så sätt att om två prov mäter samma egenskap bör korrelationen mellan proven bli stark.

Korrelationsmatrisen visar signifikanta värden på 0,05 nivå för alla de parvisa jämförelserna utom för jämförelsen mellan test A och test B. Korrelationen mellan test C och test D är hög med en korrelationskoefficient på 0,61. Mellan de övriga testerna är korrelationerna låga. Man kan alltså konstatera att de individer som har höga resultat på Test C även har det på Test D. Dessa test mäter i högre grad än de övriga ordförståelsen och orden ges i kontext i båda testen. Eleverna har alla mycket varierande kunskaper och ingen visar genomgående låga testresultat på samtliga tester. Det beror naturligtvis på att testerna har olika svårighetsgrad och mäter olika slags kunskaper.

Tabell 4.7 Interkorrelationsmatris för de olika testerna.

Korrelation	Test A	Test B	Test C	Test D
Test A	1			
Test B	0,107	1		
Test C	0,266 *	0,285 *	1	
Test D	0,290 **	0,248 *	0,608 **	1

* Korrelationen är signifikant på 0,05 nivå ($P < 0,05$)
** Korrelationen är signifikant på 0,01 nivå ($P < 0,01$)

4.6.1 Elever med svenska som modersmål och elever med annat modersmål

En jämförelse mellan elever med svenska som modersmål (antal 25), och elever med annat modersmål än svenska (antal 55), visar på skillnader mellan grupperna i de olika testerna. I test A, där orden delas i huvudsammansättningar, är skillnaderna försumbara. Det är ju inte heller ett test som har varit särskilt effektivt när det gäller att särskilja eleverna på de högre resultatintervallerna (se avsnitt 4.5.2).

I test B, självskattningen, skattar sig elever med annat modersmål i svenska på mellannivå, d.v.s. 40 – 49 poäng; något högre än eleverna med svenska som modersmål. Att fler elever med utländsk bakgrund än övriga elever ser positivt på sina kunskaper kan vara intressant att notera. Många gånger är dessa elever enligt Skolverkets ämnesrapport (2005:15) mer motiverade i skolarbetet än elever med svensk bakgrund.

I test C och D som slagits samman i tabell 4.8 har testerna delats in i poängintervall. En jämförelse mellan informanter med svenska som modersmål och övriga visar på relativt stora skillnader:

Tabell 4.8 Jämförelse i poängintervaller mellan elever med svenska som modersmål och elever med andra modersmål. Resultat i procent.

Poängintervall test C o D	Svenska som modersmål	Övriga modersmål	Total
0 – 3	0	2	1
4 – 6	16	24	21
7 – 9	20	31	28
10 – 12	20	20	20
13 – 15	44	24	30

I de lägre intervallerna 0 – 3 och 4 – 6 poäng återfinns elever med övriga modersmål i högre grad än elever med svenska som modersmål enligt tabell 4.8. Även i intervallerna 7 – 9 poäng dominerar elever med andra modersmål än svenska (31 % jämfört med 20 %). I poängintervallen 10 – 12 återfinns lika stor andel i bägge grupperna medan det i den högsta intervallen (13 – 15 poäng) återfinns betydligt fler elever med svenska som modersmål (44 % mot 24 %). Detta resultat är förväntat. Många elever som inte har svenska som modersmål har stora problem med att förstå ord och begrepp i skolans texter. SO-ämnena hör dessutom till de svåraste att tränga igenom, eftersom de innehåller många betydelsetunga och komplexa ord och begrepp som kan vara nya för eleverna och bidra till att texterna blir täta och svårgenomträngliga. Även om många av de flerspråkiga eleverna är födda i Sverige och har gått hela sin skoltid i svensk skola verkar skolspråket för dessa elever alltså vara svårt att förstå enligt denna undersökning.

4.6.2. Jämförelse mellan pojkar/män och flickor/kvinnor

Det är intressant att studera skillnad i resultat mellan pojkar/män och flickor/kvinnor för att se om det finns några allmänna tendenser. Andelen kvinnor eller män med poäng över medelvärdet i de olika testerna framgår av tabell 4.9.

Tabell 4.9 Medelvärdesresultat på de olika testen efter kön.

Test	**Medelvärdes resultat**	**Andel kvinnor med reultat över medelvärde i** %	**Andel män med reultat över medelvärde i** %
Test A	51	59	84
Test B	35	47	69
Test C	5,5	39	66
Test D	4	32	66

I Test A är medelvärdet 51 av 57 möjliga vilket visar att flertalet informanter ganska väl identifierat de olika beståndsdelarna i orden. Här har 84 % av pojkarna poäng över medelvärdet jämfört med flickornas 59 %.

I Test B, självskattningen, värderar sig pojkarna betydligt högre än flickorna; 69 % över medelvärdet och flickorna 47 %. Man kan konstatera att det verkar vara en realistisk skattning för pojkarna i förhållande till resultaten i test C och D, medan flickorna gör en högre självskattning i förhållande till resultaten.

I test C och D där orden presenteras i kontext har 66 % av männen resultat över medelvärdet mot endast 39 % respektive 32 % av kvinnorna.

I samtliga test visar pojkarna alltså ett bättre resultat än flickorna, vilket skiljer sig från normalfallet, där flickorna presterar betydligt bättre i t.ex. lästester. Detta gäller även för ord-testet på högskoleprovet (Scott 2004:25). I grundskolan och gymnasieskolan dominerar vanligtvis flickorna i de högre betygsintervallen i alla ämnen förutom i idrott och hälsa (Skolverket 2005).

4.6.3 Olika skolformer

Som tidigare nämnts går det i Test A inte att särskilja eleverna på de högre poängnivåerna. Vad man emellertid kan konstatera är, att alla informanter klarar delningen av ord relativt väl, flertalet har 50 – 57 rätt. De universitetsstuderande som bara läst svenska 1 – 2 år klarar uppgiften bättre än grundskoleeleverna som gått hela skoltiden i svensk skola, 85 % ligger på högsta poängintervall (50 – 57 rätt) medan motsvarande siffra för grundskoleeleverna är 73 %. Förmodligen har detta med läsvana och studievana att göra. De universitetsstuderande har en avslutad gymnasieutbildning i sina hemländer och har dessutom ofta också fördelen av att kunna ett västerländskt språk, vilket är till god hjälp vid förståelsen av flera testord. Test B, självskattningen, visar likartade resultat för grundskolan och gymnasieskolan där flertalet elever i båda grupperna skattar sig relativt högt (mellan 30 och 50 poäng). Även de som inte uppnår resultat på hög nivå på test C och D skattar sig högt, vilket stämmer överens med iakttagelser i andra undersökningar, nämligen att elever med uppenbara svårigheter kan överskatta sin förmåga (se avsnitt 4.2). De universitetsstuderande däremot är mer realistiska och flertalet skattar sig på lägre nivåer (20 – 50 poäng). De har ju också varit kort tid i Sverige. En studerande noterar dessutom ”att även om jag kan dela orden, så förstår jag dem inte”.

När poängen på test C och D slagits samman ser man följande resultat:

Tabel1 4.10 Resultat i poängintervaller i test C och D i procent.

Skolform	0-3 p.	4-6 p.	7-9 p.	10-12 p.	13-15 p.	Summa
Grundskola	0	17	24	29	31	100
Gymnasium	3	29	29	16	23	100
Universitet	0	31	23	15	31	100
Totalt	1	23	26	22	28	100

Som framgår av tabell 4.10 klarar grundskoleeleverna testen bäst och flertalet ligger på 10 – 15 poäng även om det också finns elever med

låga resultat (4 – 6 poäng). Även flera gymnasieelever ligger mellan 10 – 15 poäng, men alltså inte på lika höga nivåer som grundskoleeleverna. Flera gymnasieelever har inte gått hela skoltiden i svensk skola, utan fullgjort sin skolgång delvis i skolor i hemlandet, delvis i Sverige, vilket kan vara en orsak till att de inte hunnit tillägna sig ett lika stort ordförråd och flera ligger också på lägre nivåer 3 – 6 poäng. Nästan hälften av de universitetsstuderande har höga poäng (10 – 15 poäng) men 23 % uppnår medelnivå (7 – 9 poäng) och 31 % ligger på en betydligt lägre poängintervall (4 – 6) och har fortfarande stora svårigheter med förståelsen. Test C och D visar att det finns många ord och begrepp i SO-ämnena som är okända för många elever, vilket naturligtvis bidrar till att innehållet i texterna är svåra att ta till sig.

4.7 Didaktiska implikationer

Att lätt kunna genomföra olika diagnostiska ordtest vore ett önskvärt hjälpmedel och komplement i ämnesundervisningen. Det kan också ske genom datorbaserade test som eleverna kan genomföra självständigt (se kapitel 5). Test A testar indelning av ord i huvuddelar, vilket är ett viktigt sätt att uppmärksamma elever på ordens beståndsdelar och betydelsen av att kunna identifiera dessa. Det kan även göras på mer detaljerad nivå än i detta test och då också bidra till en bättre förståelse av principer för ordbildning i svenskan, vilket är en viktig kunskap vid inlärningen och förståelsen av nya ord och begrepp. Om man arbetar med ämnesorden i anknytning till undervisningen kan säkert även ämneslärarna bidra till en effektivare ordinlärning. Orden kommer då i naturligt sammanhang istället för att eleverna arbetar med lösryckta ord i isolerad språkundervisning.

Självskattningen i Test B kan kännas främmande för många elever och kanske i synnerhet för elever som inte gått i svensk skola. De kan vara ovana vid att själv bedöma sina egna kunskaper. Att bli medveten om vad man kan och inte kan är emellertid en viktig del i det egna lärandet i alla ämnen (Holmegaard 2003). Orden i testen är betydelsetunga i den meningen att de innehåller begrepp på företeelser som ofta är kulturellt och samhälleligt anknutna. Det är ju

detta som gör ett SO-ämne så komplext. Vem kan t.ex. säga att man kan ordet *kommunfullmäktige*? Det krävs ingående kunskaper för att redogöra för ett så abstrakt, kulturspecifikt och samhällsanknutet ord. Diskussioner om vad det innebär att kunna ett ord kan genom test B aktualiseras i undervisningen.

Test C och D visar sig vara effektiva test för att undersöka om man förstår ett ord med hänsyn till den omgivande kontexten. När elever stöter på ett nytt ord är det nämligen inte ovanligt att de genast tar hjälp av lexikon, utan att utnyttja ledtrådarna i kontexten. Även när eleverna blivit klara över ordets betydelse genom att slå upp det i ordboken förbiser många att kontrollera sammanhanget och att undersöka om den ordförklaring man hittat i lexikon verkar rimlig (Haastrup 1989). Att bli medveten om olika tekniker för att förstå ett ord utifrån en kontext är en viktig del i den strategiträning som undervisningen måste innehålla. Att enbart studera lösryckta ord i långa listor, något som många gånger praktiseras i skolan och vid självstudier, kan leda till att man ökar sin receptiva ordförståelse men för att kunna producera egna meningar med korrekt användning av nya ord krävs produktiva övningar med orden insatta i en relevant kontext (Nelson Wareborn 2004).

4.8 Slutsatser

Den här redovisade undersökningen visar att långa ord med flera stavelser i SO-ämnena är svåra att förstå för många elever (detta gäller både elever med svenska som första- och andraspråk) och kan hindra läsförståelsen. Det handlar ofta om betydelsetunga substantiv som kräver sociokulturell förståelse. De ord som förefaller svårast för andraspråkseleverna är tämligen abstrakta och allmänna läroboksord såsom *inflytande, rörelse* och *företräda*. En extra svårighet tycks polysema ord utgöra. Här kan det sammansatta ordet innehålla en del som har en betydligt vanligare betydelse än den som avses i sammansättningen som t. ex *rörelse, utskott* och förstavelsen *ställ-*. Det är dessutom ord som är sällsynta i OrdiL-materialet och som därför kan vara extra svåra.
Att kunna dela in orden i huvudled (Test A), vilket flertalet elever

klarade av väl, är naturligtvis ingen garanti för att man förstår orden. Det kan ändå vara en mycket god hjälp när man ska försöka förstå nya sammansatta ord.

Självskattningen (Test B) är ett bra komplement vid testning av ordkunskap (Oscarsson 1999). Här får man en god uppfattning om vilka ord som anses svåra av informanterna. Genom att kategorisera dessa ord, vilket inte gjorts i denna studie, kan man få empiriskt underlag för ett relevant urval av ord för kommande tester. Graden av korrelation mellan självskattningen i Test B och matchningstestet Test C är låg och detta gäller även mellan Test B och flervalstestet Test D.

Elever med svenska som modersmål klarar testerna C och D klart bättre än övriga elever, vilket är ett förväntat resultat. Indelning av ord (Test A) visar däremot inga stora skillnader mellan grupperna. Detta test skulle emellertid kunna utvecklas och förändras för att ge ett bättre utslag, bl.a. genom att tiden för genomförandet kortades ned och takeffekten minimerades. Elever med få poäng på detta test borde också undersökas närmare. Eventuellt kan detta test vara effektivt för att skilja ut elever med olika behov, men här krävs en större studie för mer tillförlitliga resultat.

Om man ser på resultaten i de olika skolformerna, ser man att grundskoleeleverna klarar testen bäst, men att spridningen är mycket stor. Detta kan visa en tendens som även uppmärksammats i andra tester (PISA 2003), nämligen att kunskapsklyftan mellan mycket starka och svaga elever har vidgats. Det kan också noteras att flera universitetsstuderande, trots att de endast varit några år i Sverige har klarat av testerna relativt väl. Man kan förmoda att det hänger ihop med deras skolbakgrund och kanske även med det faktum att flera begrepp är bekanta för dessa informanter på modersmålet och därigenom kan överföras till svenska. De behärskar även annat västerländskt språk, vilket underlättar förståelsen av flera ord med latinskt och grekiskt ursprung.

I testen uppnår pojkarna högre poäng än flickorna, vilket inte överensstämmer med resultaten från andra ordtest, där flickor ofta

får betydligt högre poäng. De fyra testen visar också på vissa inbördes samband (se tabell 4.7). Analyserna visar på signifikanta samband i fem av de sex genomförda korrelationsanalyserna. Sambandet är svagt mellan Test A, Test C och Test D, samt mellan Test B, Test C och Test D. Det starkaste sambandet finns mellan Test C och Test D där korrelation är 0,61. Sambandet mellan Test A och Test B är inte signifikant. Här skulle man behöva upprepa testen i större skala och i mer enhetligt sammansatta grupper för att komma fram till mer reliabla resultat.

Referenser

Adams, M. J. 1990. *Beginning to read: Thinking and learning about print*. Cambridge: The MIT Press.

Bachman, L. & Palmer, A.S. 1989. The construct validation of selfratings of communicative language ability. *Language Testing*, 6, s. 14 – 25.

Björnson, C.H. 1968. *Läsbarhet*. Stockholm: Liber.

Brown, J.D. 1997. An EFL readability index. *University of Hawaii Working papers in English as a Second Language*, s. 85 – 119.

Carrell, P. L. 1987. *Readability in ESL*. 4, s. 21 – 40.

Collier, V. & Thomas, W. 2002. *A National Study of School Effectiveness for Language minority Students, Long-Term Academic Achievement*. Final Report. CREDE: Centre for Research on Education, Diversity and Excellence. Washington. http//www.crede.ucsc.edu/research/llaa/1.1 final.html. 2006-11-22

Conners, F.A. & Olsson, R. K. 1990. Reading comprehension in dyslexic and normal readers: A component skill analysis. I Balota, D., A. Flores d'Arcais, G. B. & Ryner, K.
(red.). *Comprehension processes in reading,* s. 557 – 579. Hillsdale, NJ: Erlbaum Associates.

Craik, F.I. & Lockhart, R. S. 1972. Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior* 1, s. 671 – 684.

von Elek, T. 1981. *Självbedömning av färdigheter i svenska som andraspråk*. Delrapport från utvecklingsprojekt, Working Papers from the Language Teaching Research Center, No30. Department of Education, Göteborgs University, Sweden.

von Elek, T. 1985. A test of Swedish as a second language. I Y. P. Lee et al. (red.). *New Directions in Language Testing*. Oxford: Pergamon Press.

Haastrup, K. 1989. The Learner as Word Processor. I: Nation, P. &
Carter, R. (red.). *Vocabulary Acquisition*. AILA Review 6.), s. 34 – 46.

Harrison, C. 1980. *Readability in the Classroom*. Cambridge: Cambridge University Press.

Holmegaard, M. 1999. *Språkmedvetenhet och ordinlärning. Lärare och inlärare reflekterar kring en betydelsefältsövning i svenska som andraspråk.* Göteborg: Acta Universitatis Gothoburgensis.

Holmegaard, M. 2003. *Ett försök med självbedömning under lärarutbildningen.* UFL-rapport 2003:1. Göteborgs universitet: Utbildnings- och forskningsnämnden för lärarutbildning.
http://www.ufl.gu.se/digitalAssets/719597 Rapport 200301 pdf 06-11-01.

Jacobsen, C. 199*3. Ordkedjor. Manual.* Stockholm: Psykologiförlaget.

Johansson, M-G. 1999. *Fyra korta lästest för snabb och enkel bedömning av läsfärdighet.* MG-kedjor, Handbok. Frösön: MG Läs- och Skrivkonsult AB.

Lewkowicz, J. A. & Moon, J. 1985. Evaluation, a way of involving the learner. In J. C. Alderson (red.). Lancaster Practical Papers in English Language Education, vol. 6: *Evaluation,* s. 45 – 80. Oxford: Pergamont Press.

Lundberg, I. 1984. *Språk och läsning.* Malmö: Liber.

Meara, P. 1996. The Classical Research in L2 Vocabulary Acquisition. I Anderman, Gunilla & Rogers, Margaret (eds.). *Words, Words, Words.* The Translator and The Language
Learner. Clevedon: Multilingual Matters, s. 27 – 41.

Meara, P. & Jones, G. 1988. Vocabulary size as a placement indicator. I P. Grunwell (ed.). *Applied Linguistics Society* (S.80-87) London: Centre for Information on Language Teaching and Research.

Melander, B. 2004. Läsebokssvenska, bruksprosa och begreppslighet. En översikt över svensk språkforskniing kring läroböcker. I Strömquist, S. (red.). *Läroboksspråk.* ORD OCH STIL Språksamfundets skrifter 26. Uppsala: Hallgren och Fallgren Studieförlag AB, s. 12 – 46.

Melka, F. 1997. Receptive vs. productive aspects of vocabulary. I *Schmit, N. and Mc Carthy, M.* (Eds.). *Vocabulary: Description, Acquisition and Pedagogy.* Cambridge: Cambridge
University Press, s 84 – 102.

Miller Guron, L. 1999. *Wordchains. A word reading test for all ages.* Windsor: NFER-NELSON.

Nelson Wareborn, M. 2004. *Minst 100 nya ord i veckan ska läras in. Vem har motivation och uthållighet för det? Andraspråksinlärares strategianvänd-*

ning i ordinlärning. IPD-rapport 2004:4. Göteborgs universitet. Institutionen för Pedagogik och Didaktik.
Oscarsson, M. 1999. Estimating Language Ability by Self-Assessment: A review of Some of the issues. I *Papers on Language, Learning, Teaching, Assessment.* IPD-rapport nr 1999:02. Göteborgs universitet. Institutionen för pedagogik och didaktik.
PISA 2000, 2003. Programme for International Student Assessment. *Svenska emotonåringars läsförmåga och kunnande i matematik och naturvetenskap i ett internationellt perspektiv.* Stockholm: Skolverket.
Scott, S. 2004.*Ordförståelse.* En litteraturstudie med anknytning till högskoleprovets ORD-prov. BVM nr 2. Umeå universitet. http://www.umu.se/edmeas/pblikationer/pdf BVM%20nr%202.pdf 2006-11-01.
Read, J. 2000. *Assessing Vocabulary.* Cambridge: Cambridge University Press.
Skolverket 2005. Skolverket. Nationella utvärderingen av grundskolan 2003. Ämnesrapport till RAPPORT 251, 2005. *Svenska och svenska som andraspråk årskurs 9.*
Skolverket, 2005-23-09, http://www.skolverket.se
Shrauger, J.S. & Osberg, T. M. 1981. The relative accuracy of self-predictions and judgments by others in psychological assessment. *Psychological Bulletin,* 90, 2, s. 322 – 351.
Strong-Krause, D. 1997. *How effective is self-assessment for ESL placement?* Paper presented at TESOL 1997, Orlando, Florida, mars 1997.
Svensk skolordlista. 2004. Stockholm: Förlag Norstedt.
Taube, K. 2000. *Hur läser Stockholms elever med annat modersmål än svenska?* Dyslexi 2. Svenska Dyslexiföreningen, Stockholm.
Wahlgren, L.2005. *SPSS steg för steg.* Lund: Studentlitteratur.

Bilagor

Bilaga 1a Enkät
Bilaga 1b Bakgrundsfakta
Bilaga 2 Test A
Bilaga 3 Test B
Bilaga 4 Test C
Bilaga 5 Test D

Bilagor

Bilaga 1a Enkät

Tack för att du hjälper till att prova ut tester i några läroböcker! Vi är en forskargrupp vid Göteborgs universitet, som vill veta mer om svårigheter i olika läromedel och det är därför mycket viktigt för oss att veta dina synpunkter och kunskaper.

Följ din lärares instruktioner! Du är helt anonym, men senare kan du få diskutera testorden med din lärare.

Hälsningar Margareta Holmegaard

Var snäll och fyll först i följande bakgrundsuppgifter:

Födelseår…… I vilket land?.....Kom till Sverige år……?

Kvinna ☐ man ☐

Modersmål……

Vilket/a språk talar du med dina föräldrar?.....

Vilket/a språk talar du med dina syskon?.....

Skolgång i Sverige……Antal år……

Skolgång i annat land……..Ange vilket land……

Deltar i svenska som andraspråksundervisning Ja ☐ Nej ☐ Vet ej ☐

Deltar i svenskundervisning Ja ☐ Nej ☐ Vet ej ☐

Deltar i modersmålsundervisning Ja ☐ Nej ☐

Bilaga 1b Bakgrundsuppgifter om informanterna

	Totalt	Modersmål		Skolform			Kön	
		Svenska	Annat spec.	Grund-skola	Gymna-sium	Univer-sitet	Kvinnor	Män
Antal informanter	87	25	55 (Bortfall 7)	43	31	13	55	32
Kön								
Kvinnor, antal (%)	55 (63)	13 (52)	37 (67)	20 (47)	27 (87)	8 (62)	55 (100)	0 (0)
Män, antal (%)	32 (37)	12 (48)	18 (33)	23 (53)	4 (13)	5 (38)	0 (0)	32 (100)
Födelseår								
1984 – 1987, antal (%)	24 (28)	1 (4)	23 (42)	0	11 (35)	13 (100)	17 (31)	7 (22)
1988, antal (%)	17 (20)	6 (24)	11 (20)	0	17 (55)	0	16 (29)	1 (3)
1989, antal (%)	41 (47)	16 (64)	18 (33)	38 (88)	3 (10)	0	19 (35)	22 (69)
1990, antal (%)	5 (6)	2 (8)	3 (5)	5 (12)	0	0	3 (5)	2 (6)
Tid i Sverige och skoltider								
Tid i Sverige, median (variationsvidd)	15 (1,5-19)	16 (15-19)	12,5 (1,5-17)	15 (5-16)	13,5 (3-19)	1,5 (1,5-4)	15 (1,5-17)	15 (1,5-19)
Skoltid i Sverige, median (variationsvidd)	9 (1-12)	9 (5-12)	9 (1-12)	9 (5-10)	10 (3-12)	1,5 (1-2,5)	9 (1-12)	9 (1,5-12)
Skoltid i annat land, median (variationsvidd)	0 (0-14)	0	0 (0-14)	0 (0-4)	0 (0-9)	12 (11-14)	0 (0-14)	0 (0-13)
Total skoltid, median (variationsvidd)	9 (5-15,5)	9 (5-12)	10 (5-15,5)	9 (5-10)	10 (5-13)	13,5 (12-15,5)	10 (5-15,5)	9 (5-15)
Skolform								
Grundskola, antal (%)	43 (49)	18 (72)	18 (33)	43 (100)	0 (0)	0 (0)	20 (36)	23 (72)
Gymnasium, antal (%)	31 (36)	7 (28)	24 (44)	0 (0)	31 (100)	0 (0)	27 (49)	4 (13)
Universitet, antal (%)	13 (15)	0	13 (24)	0 (0)	0 (0)	13 (100)	8 (15)	5 (16)
Ämnesdeltagande								
Svenska som andraspråk, antal (%)	45 (52)	1 (4)	44 (80)	11 (26)	21 (68)	13 (100)	30 (55)	15 (47)
Svenska som modersmål antal (%)	48 (55)	24 (96)	18 (33)	37 (86)	11 (35)	0 (0)	26 (47)	22 (69)
Både svenska och sva, antal %	7 (8)	0	7 (13)	6 (14)	1 (3)	0 (0)	5 (9)	2 (6)
Modersmålsundervisning, antal (%)	14 (16)	1 (14)	13 (24)	6 (14)	8 (26)	0 (0)	8 (15)	6 (19)
Modersmål								
Svenska, antal (%)	25 (29)	25 (100)	0	18 (42)	7 (23)	0 (0)	13 (24)	12 (38)
Arabiska, antal (%)	8 (9)	0	8 (15)	3 (7)	2 (6)	3 (23)	5 (9)	3 (9)
Persiska, antal (%)	5 (6)	0	5 (9)	2 (5)	1 (3)	2 (15)	3 (5)	2 (6)
Spanska, antal (%)	5 (6)	0	5 (9)	1 (2)	2 (6)	2 (15)	3 (5)	2 (6)
Övriga specificerade modersmål, antal (%)*	37 (43)	0	37 (67)	12 (28)	19 (61)	6 (46)	26 (47)	11 (34)
Ej uppgivit något modersmål, antal (%)	7 (8)	0	0	7 (16)	0 (0)	0 (0)	5 (9)	2 (6)

Bilaga 2

Test A
Du skall under 10 minuter arbeta med orden i testen som valts ut ur olika läroböcker i SO-ämnena. Dela orden i två delar efter deras huvudsammansättningar som i exemplet:
Serie/tidning, tidnings/läsning
Börja arbeta med orden när din lärare ger dig klartecken och markera så många ord du hinner.

Högmässogudstjänst	Yttrandefrihetsgrundlagen	Gästgivaregård
Frälsningssoldat	Tryckfrihetsförordningen	Budgetproposition
Ungdomsverksamhet	Arbetsmarknadsutskottet	Parlamentsledamöter
Musikgudstjänst	Landsbygdsbefolkning	Kvinnodemonstration
Sammanslagning	Riksdagsledamöterna	Nationalförsamlingen
Vigselförrättare	Åldersfördelningen	Samhällsklasser
Ställföreträdare	Utbildningsnämnden	Produktionsmedlen
Nattvardsfirande	Trafikföreskrifter	Förhandlingsvilja
Motreformation	Styrkeförhållanden	Stenkolsgruvor
Valsensation	Samhällsinflytande	Missförhållanden
Begravningsceremoni	Oppositionspartier	Medellivslängd
Reformationsrörelse	Regeringskris	Massarbetslöshet
Gudstjänstbesökare	Kommunfullmäktige	Majoritetsbeslut
Konfirmationsakt	Fackförbundsordförande	Järnvägsknutar
Samfundsgränser	Tillverkningsindustri	Exportindustri
Utrikespolitik	Justitiedepartement	Partitillhörighet
Församlingsbor	Fackföreningsrörelse	Ostindiekompaniet
Munkkloster	Utbildningsutskott	Kommunsammanslagning
Pastorsexpedition	Historieskrivning	Representanthuset

Bilaga 3

Test B
Studera nu igen orden som du delat. Skriv **Ja, Nej** eller **?** i rutan jämte orden enligt följande alternativ:
Ja = Jag vet säkert vad ordet betyder
Nej = Jag vet inte alls vad ordet betyder
? = Jag är osäker på vad ordet betyder

Högmässogudstjänst ☐	Yttrandefrihetsgrundlagen ☐	Gästgivaregård ☐
Frälsningssoldat ☐	Tryckfrihetsförordningen ☐	Budgetproposition ☐
Ungdomsverksamhet ☐	Arbetsmarknadsutskottet ☐	Parlamentsledamöter ☐
Musikgudstjänst ☐	Landsbygdsbefolkning ☐	Kvinnodemonstration ☐
Sammanslagning ☐	Riksdagsledamöterna ☐	Nationalförsamlingen ☐
Vigselförrättare ☐	Åldersfördelningen ☐	Samhällsklasser ☐
Ställföreträdare ☐	Utbildningsnämnden ☐	Produktionsmedlen ☐
Nattvardsfirande ☐	Trafikföreskrifter ☐	Förhandlingsvilja ☐
Motreformation ☐	Styrkeförhållanden ☐	Stenkolsgruvor ☐
Valsensation ☐	Samhällsinflytande ☐	Missförhållanden ☐
Begravningsceremoni ☐	Oppositionspartier ☐	Medellivslängd ☐
Reformationsrörelse ☐	Regeringskris ☐	Massarbetslöshet ☐
Gudstjänstbesökare ☐	Kommunfullmäktige ☐	Majoritetsbeslut ☐
Konfirmationsakt ☐	Fackförbundsordförande ☐	Järnvägsknutar ☐
Samfundsgränser ☐	Tillverkningsindustri ☐	Exportindustri ☐
Utrikespolitik ☐	Justitiedepartement ☐	Partitillhörighet ☐
Församlingsbor ☐	Fackföreningsrörelse ☐	Ostindiekompaniet ☐
Munkkloster ☐	Utbildningsutskott ☐	Kommunsammanslagning
Pastorsexpedition ☐	Historieskrivning ☐	Representanthuset ☐

Bilaga 4

Test C

Den katolska motreformen fick stor betydelse.	10
Exempel på stora organisationer är fackföreningsrörelsen, idrottsrörelsen och miljörörelsen.	
Frågor som tar upp undervisning behandlas i utbildningsutskottet.	
Alla visade prov på förhandlingsvilja.	
En trafikföreskrift kan behandla frågor som berör hastighet på en vägsträcka och omkörningar.	
Det kom många gudstjänstbesökare.	
Riksdagsledamöterna ställde upp.	
Majoritetsbeslut måste alltid följas.	
Ni får agera som ställföreträdare.	
Det finns flera oppositionspartier i landet.	

1 En form som inte är som i beskrivningen

2 Instruktioner som skall följas av bilister.

3 De politiska grupper som inte är med i regeringen.

4 Personer som ersätter någon (får ta en annans plats).

5 De flesta människor är med och bestämmer något.

6 Kommer ofta till kyrkan på söndagar.

7 Diskuterar skolfrågor innan riksdagen fattar beslut.

8 Grupp av politiker som folket valt och som är med om att styra landet.

9 Diskuterar löner och hjälper anställda som inte kommer överens med sin arbetsgivare

10 De ville inte ha någon religiös förändring

11 Många tjänar sitt uppehälle genom arbete i kyrkan

12 Man diskuterar för att komma överens

13 Regel för fotgängare

14 Skrifter som handlar om bilar

Test D

Välj en mening som passar som förklaring. **Markera med en ring** kring den siffra som korrekt klargör det understrukna ordet.

a) De hade skaffat en ny lokal för justitiedepartementet.
1. har hand om rättsväsendet
2. har kunskaper inom flera vetenskaper
3. planerar våra vägar
4. utbildar personal

b) Kommunfullmäktiges sammanträden är öppna för allmänheten.
Politiker som
1. fattar de högsta besluten i en kommun
2. har hand om sjukvården i en kommun
3. är mäktigast i landet och kommunen
4. har hand om landets och kommunens lagar

c) De hade själva valt ut vigselförrättaren.
En person som:
1. tjänstgör vid ceremoni då två personer gifter sig
2. tjänstgör vid begravningar
3. arbetar vid examinationer
4. arbetar med släktfrågor

d) Det behövs ökat samhällsinflytande inom de stora företagen
1. mer pengar måste flyta in
2. mer information till företagen
3. Riksdag och regering måste bestämma mer inom företagen
4. Företagen måste själva bestämma mer

e) Tillverkningsindustrin ställer stora krav på goda transporter
1. sysslar med information
2. arbetar fram färdiga föremål
3. gör nya vägar
4. arbetar med kommunikationer

f) Massarbetslösheten bredde ut sig i hela landet
1. brist på arbete bland industriarbetare
2. stor mängd människor är utan arbete
3. människor inom massaindustrin är utan arbete
4. arbetsbrist för människor inom massmedia

5. Webb-versioner av diagnostiska prov

Sofie Johansson Kokkinakis

I detta avsnitt beskrivs en webb-baserad version av diagnostiska prov som ursprungligen utvecklats av Margareta Holmegaard (MH), se kapitel 4. Den datorbaserade webb-versionen har utvecklats av Sofie Johansson Kokkinakis (SJK). Beskrivning av webb-versionen i texten i detta kapitel står SJK för. Nedan anges exempel från provet, samt diskussioner om fördelar och nackdelar med denna typ av prov.

Fördelen med att använda en datorbaserad webb-version av ett prov är att provsvar direkt kan skickas elektroniskt till den person som är ansvarig för proven. Varje fråga i proven motsvarar en variabel med ett i förväg angivet värde eller ett okänt värde. Då formuläret (se 5.1) är ifyllt kan eleven klicka på en knapp för att skicka provet. Tid kommer att registreras och i vissa fall även styra provtagningen. Processen underlättar insamling och bearbetning av provsvaren.

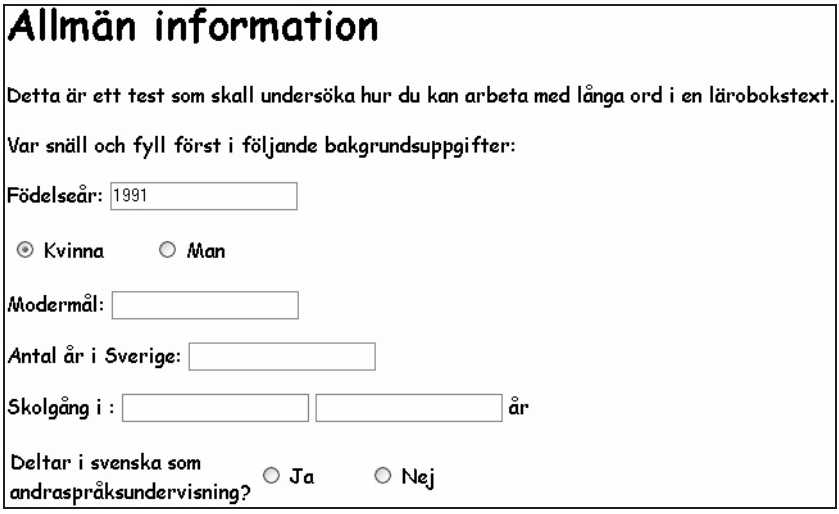
För att säkerställa elevernas identitet får eleverna lämpligen en kod från läraren som används vid inloggning på webb-sidan för provet.

En skillnad vid användning av webb-baserade prov är att eleverna förmodligen inte kommer att kunna utföra testerna samtidigt då det kräver ett flertal datorer. Det kan då vara en fördel att proven inte blir för långa för att undvika att uppgifter om proven diskuteras mellan eleverna.

I avsnittet 5.1 kommer exempel på hur ett prov kan se ut illustreras samt information som skickas vidare till ansvarig för provet. I detta projekt har statistikprogrammet SPSS (2001) använts för att bearbeta provsvar.

5.1 Utformning av webb-baserade diagnostiska prov

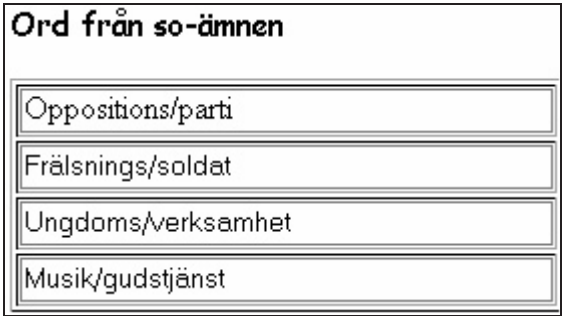
Figur 5.1 ger eleven möjlighet att fylla i bakgrundsuppgifter. I fall då flera prov ska utföras bör ett sådant formulär endast ifyllas en gång.



Figur 5.1 Bakgrundsuppgifter.

Vi avser i projektet att följa etiska regler vad gäller registrering av bakgrundsinformation om en elev samt PUL (Personuppgiftslagen). Därför kommer endast uppgifter som anses vara intressanta ur

forskningssynpunkt att bevaras. En elev som gör ett prov bör ha ett identitetsnummer vilket blir unikt för det provet. I ett formulär som i figur 5.1 sparas information som beskriver ett unikt identitetsnummer, födelseår och kön etc. Här följer några exempel på hur proven kan utformas. I figur 5.2 får eleven dela ord efter deras huvudsammansättningar genom att dela av ordet med ett snedstreck ”/”.



Figur 5.2 Dela orden i två delar efter deras huvudsammansättningar.

Vid analys av sparad information om uppgiften i figur 5.2 så kommer svar om eleven valt rätt eller fel alternativ att framgå. Vid felaktiga svar kan ev. det felaktiga svaret sparas för vidare analys.

Det aktuella provet testar också om eleven tror sig känna till orden, se figur 5.3.

Figur 5.3 Frågor om eleverna vet vad orden betyder.

I övningen som beskrivs i figur 5.4 undersöks om eleven kan hitta matchande beskrivningar till understrukna ord. Resultat kan redovisas i form av rätt eller fel.

Det finns flera oppositionspartier i landet.	3	1 En form som inte är som i beskrivningen
Exempel på stora organisationer är fackföreningsrörelsen, idrottsrörelsen och miljörörelsen.		2 Instruktioner som skall följas av bilister.
Frågor som tar upp undervisning behandlas i utbildningsutskottet.	5	3 De politiska grupper som inte är med i regeringen.
Alla visade prov på förhandlingsvilja.		4 Personer som ersätter någon (får ta en annans plats)
En trafikföreskrift kan behandla frågor som berör hastighet på en vägsträcka och omkörningar.	2	5 Diskuterar skolfrågor innan riksdagen fattar beslut

Figur 5.4 Matcha understrukna ord med numrerade meningar.

Det sista testet i figur 5.5 går ut på att matcha ett understruket ord med en förklaring. Eleven får då markera det svar som hon/han tror är rätt.

	a)	De hade skaffat en ny lokal för justitiedepartementet.
☑	1	har hand om rättsväsendet
☐	2	har kunskaper inom flera vetenskaper
☐	3	planerar våra vägar
☐	4	utbildar personal

Figur 5.5 Välj en mening som passar som förklaring till understruket ord.

5.2 Slutsatser om Webb-baserade versioner av diagnostiska prov

Denna form av datorbaserade prov kommer att bli mycket tidsbesparande jämfört med att utföra proven med penna och papper. En provansvarig behöver inte vara på plats när eleverna har blivit vana vid proven. Genom att kopiera in resultatdata som enkelt skickas elektroniskt via provformulären undviks tidsödande inmatning av resultaten. På detta sätt underlättas även statistiska beräkningar baserade

på resultat. Det kan även bli möjligt att logga elevens datorrelaterade aktiviteter samt tidsåtgång i samband med utförandet av provet. Ytterligare en fördel är också att man undviker en svårighet som ibland uppstår med att tolka provsvar då dessa är handskrivna.

För att utföra proven krävs mycket liten datorvana, men samtidigt är det viktigt att identiteter skapas för varje elev för att undvika fusk. Man kan även tänka sig att proven utformas på olika sätt och anpassas beroende på om det är för en forskare som proven görs eller om det är en lärare som använder proven för att diagnostisera eleverna.

Problem som kan tänkas uppstå i samband med datorbaserade prov är t.ex. att en lärare eller provansvarig inte får samma kontroll över utförandet av provet om det inte utförs samtidigt av alla elever, möjligtvis fungerar det bra om det finns datorer i klassrummet. Men samtal om proven mellan elever är alltid en risk då provtagningen pågår under en längre period än om alla utför provet samtidigt. En annan risk med datorbaserade prov är att elever med mindre datorvana känner sig stressade av datormiljön eller hindrade då de inte förstår hur proven ska utföras. Det är därför viktigt att proven utformas med så enkelt och lättförståeligt gränssnitt som möjligt och att i samband med utförandet också undersöka om eleverna har använt datorer i skolan i sin skrivprocess.

ROSA. *Denna rapport är nummer 8 i den oregelbundet utkommande serien ROSA (Rapport Om Svenska som Andraspråk) som utkommer vid Institutet för svenska som andraspråk, Institutionen för svenska språket, Göteborgs universitet.*

ORDFÖRRÅDET I SKOLANS LÄROMEDEL kan vara en stor stötesten för många elever i grundskolan - och kanske i synnerhet för flerspråkiga elever. Att tillägna sig de ord och fraser som krävs för att förstå det som står i läroböckerna i olika ämnen kan lätt te sig som en oöverstiglig uppgift. Behovet av en effektiv och systematisk språkundervisning är därför stort.

I denna rapport redovisas arbetet i ett projekt som syftar till att kartlägga ordförrådet i vanliga läroböcker i NO, SO och matematik för grundskolans senare år. En central del i projektet är uppbyggnaden av en läromedelskorpus, där läromedelstexter i NO- och SO-ämnen samt i matematik lagts in.

En sådan inventering kan bidra med viktig kunskap om lexikala aspekter av ett skolbaserat och skolrelevant språk. Det är information som bl.a. kan läggas till grund för ett mer systematiskt, effektivt och ämnesövergripande pedagogiskt arbete med utvecklingen av ett skolbaserat ordförråd. Läromedelskorpusen ger också ett utmärkt underlag för konstruktion av olika typer av diagnostiska bedömningsinstrument som kan användas för att mer systematiskt följa utvecklingen av elevernas receptiva och produktiva ordförråd ur såväl kvantitava som kvalitativa aspekter.

I denna rapport redovisas det arbete som genomförts i projektet fram till och med juni 2006.

ISSN 1403–1353