NORDISKE STUDIER I LEKSIKOGRAFI

Titel: Från *aspekt* till *övergripande* – en ordlista över svensk akademisk vokabulär

Forfatter: Judy Ribeck, Håkan Jansson & Emma Sköldberg

Kilde: Nordiske Studier i Leksikografi 12, 2013, s. 370-384
Rapport fra Konferanse om leksikografi i Norden, Oslo 13.-16. august 2013

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

Søgbarhed
Artiklerne i de ældre Nordiske studier i leksikografi (1-5) er skannet og OCR-behandlet. OCR står for ’optical character recognition’ og kan ved tegngenkendelse konvertere et billede til tekst. Dermed kan man søge i teksten. Imidlertid kan der opstå fejl i tegngenkendelsen, og når man søger på fx navne, skal man være forberedt på at søgningen ikke er 100 % pålidelig.

Från *aspekt* till *övergripande* – en ordlista över svensk akademisk vokabulär

Judy Ribeck, Håkan Jansson & Emma Sköldberg

This report describes a project to develop an academic word list for Swedish. The resulting word list is published at <http://spraakbanken.gu.se/ao/>. It comprises 655 headwords, extracted from a 25 million word corpus of Swedish academic texts. Both the word list and the corpus are openly accessible through Språkbanken's lexical and corpus infrastructures.

1. Inledning

Betydelsen av att tillägna sig ett adekvat ordförråd för akademisk verksamhet har sedan en tid fått ökad uppmärksamhet. Till en början gällde intresset i huvudsak akademisk engelska, men nu växer insikten om behoven även för andra språk. Mot denna bakgrund har forskare sedan 1970-talet tagit fram olika *akademiska ordlistor*.

Vi presenterar här vårt arbete med att framställa en *svensk akademisk ordlista*. Syftet med denna är att stödja, i första hand, andraspråksinlärare på mer avancerad nivå, men även modersmålsstudenter som är ovana vid det akademiska språket. Listan riktar sig också till dem som väljer att skriva på svenska i stället för på engelska inom ramen för sina högre studier.

Ordlistan är främst tänkt att användas vid produktion av akademisk text, men tack vare att uppslagsorden försetts med svenska betydelseangivelser och engelska översättningsekvivalenter, kan den även vara till hjälp vid reception. Dessutom kan vårt arbete tjäna som underlag för utveckling av framtida ordtester och undervisningsmaterial samt bidra till att dokumentera det svenska akademiska ordförrådet.

Ända sedan projektets början har vi kontinuerligt redogjort för dess fortgång i olika sammanhang. Vi hänvisar därför den som vill följa våra metodologiska överväganden till tidigare publikationer.[1]

Denna artikel inleds med en kort forskningsbakgrund, varefter vi beskriver vårt korpusmaterial, vår metod för urval av uppslagsord samt vår resulterande ordlistas innehåll och utformning. Texten avslutas med några ord om framtida utvecklingsbehov.

2. Tidigare akademiska ordlistor

I engelskspråkig litteratur står den övergripande termen *aca-*

1. Se Sköldberg & Johansson Kokkinakis (2012) för en allmän projektpresentation, Jansson et al. (2012) om insamling av akademiska texter, Johansson Kokkinakis et al. (2012) om nordiskt samarbete och Carlund et al. (2012) om CALL. Vi vill dock understryka att dessa arbeten behandlar en tidigare version av ordlistan, som byggde på ett mindre omfattande korpusmaterial.

demic vocabulary för en typ av ord som förekommer frekvent i löpande akademisk text från flera olika discipliner (se Paquot 2010:17–21 för en termutredning). Med början på tidigt 1970-tal har en rad olika engelska akademiska ordlistor framställts, för att möta behoven hos studenter på avancerad nivå. Den som har haft störst genomslag är *The Academic Word List* (AWL, Coxhead 2000 m.fl.).

AWL omfattar sammanlagt 570 *ordfamiljer*[2], fördelade över 10 frekvensbaserade dellistor. Listan är baserad på en korpus på 3,5 miljoner ord som utgörs av akademiska artiklar och kurslitteratur. Korpusen är indelad i fyra discipliner, vilka, var och en, innehåller sju ämnesområden. Förutom de allmänna, ovan nämnda kriterierna för akademiska ord, gäller för ordfamiljerna i AWL att dess medlemmar inte får tillhöra de 2000 vanligaste orden i språket.

Under det senaste decenniet har AWL använts flitigt i språkundervisning, ordkunskapstest och läroböcker, samt som forskningsunderlag (Coxhead 2011). Listan har dock inte undgått kritik. I huvudsak går kritiken ut på följande:

1) Användningen av ordfamiljer snedvrider ordurvalet. Ord, som inte av egen kraft uppfyller kriterierna, «räddas» genom avledningar med annan betydelse. Vidare tas ingen hänsyn till homografi och polysemi; gemensamt etymologiskt ursprung behöver inte innebära att ordstammen har samma betydelse i besläktade ord (Wang Ming-Tzu & Nation 2004).

2) Indelningen och urvalet av korpustexter är godtycklig (Hyland & Tse 2007).

2. Idén att använda s.k. *ordfamiljer* vid ordinlärning presenterades i Bauer & Nation (1993). Med en ordfamilj avses en ordstam och alla dess vanliga böjningar och avledningar, t.ex. *react*, *reacting*, *reaction*, *reactionary*, *reactive*, *unreactive*, *reactivate*, *reactor*.

3) Föresatsen att exkludera vardagliga ord har ifrågasatts, då vissa ord ur basordförrådet anses ha speciella akademiska funktioner (jfr Paquot 2010, Gardner & Davies 2013). Här utmanas följaktligen hela idén om att ordförrådet kan delas in i diskreta stycken, som kan studeras var för sig (jfr Nation 2001). Det är dessutom problematiskt att, som Coxhead, grunda exkluderingen av ord på en jämförelse med en föråldrad ordlista som *General Service List* från 1953.

Olika forskare är alltså inte överens om hur man bäst beskriver och identifierar akademiskt ordförråd. Detta har, på senare år, resulterat i åtminstone två seriösa utmanare till AWL: *the Academic Keyword List* (AKL, Paquot 2010) och *the Academic Vocabulary List* (AVL, Gardner & Davies 2013). Båda dessa ordlistor räknar med lemman, i stället för ordfamiljer. Båda kräver också att orden är jämnt fördelade över akademiska texter, samt utmärkande för just denna texttyp (s.k. *nyckelord*). Det senare kravet uppfylls genom jämförelser med referenskorpusar.

AKL innehåller 930 akademiska nyckelord, som extraherats ur en korpus på 3 miljoner ord, fördelade över akademisk prosa och studentuppsatser. Texterna kommer från fem olika discipliner. I materialet ingår även en skönlitterär referenskorpus.

AVL består av 3000 ord, framtagna ur ett korpusmaterial på över 120 miljoner ord. Texterna utgörs av tidskriftsartiklar från nio olika discipliner. Som jämförelsematerial används referenskorpusar med nyhetstexter och skönlitteratur. Förutom det stora empiriska underlaget och listans omfattning, är AVL även unik i det att orden inte får förekomma «oväntat» mycket i någon eller några få discipliner.

3. Metod och material

Den metod som vi använt för att ta fram kandidater till den svenska akademiska ordlistan är tydligt inspirerad av tidigare försök att extrahera akademiska ord ur en korpus med akademiska texter. Det första steget består således i att, på bästa sätt, sätta samman en korpus som är representativ för svenskt akademiskt skriftspråk (se 3.1). Nästa steg är att i denna korpus identifiera akademiska ord (se 3.2).

3.1. SveAk

SveAk – Svensk akademisk korpus – består av sammanlagt 25,4 miljoner ord från avhandlingar och tidskriftsartiklar som publicerats 1997–2012. Korpusen är fritt tillgänglig via korpusinfrastrukturen, *Korp*, i Språkbanken (se Borin et al. 2012b för närmare beskrivning). Vid textinsamlingen har vi utgått från den nationella databasen *SwePub*, som listar alla publikationer från svenska universitet och högskolor enligt en internationell standard.[3]

3. Se Jansson et al. (2012:958) och där angivna referenser för närmare detaljer.

Humaniora	**Ord**[4]	**Samhällsvetenskap**	**Ord**
Etnologi	1 669	Ekonomi/näringsliv	1 886
Filosofi	853	Juridik	683
Historia	2 704	Medie-/kommunikations-vetenskap	1 131
Konst	1 650	Psykologi	340
Litteratur-vetenskap	2 359	Social/ekonomisk geografi	1 621
Religion	2 957	Sociologi	1 838
Språk-vetenskap	2 287	Statsvetenskap	1 557
		Utbildningsvetenskap	1 827
Totalt	**14 479**		**10 883**

Tabell 1: SveAk:s sammansättning.

För att summera de tankar som legat till grund för SveAk:s sammansättning (se tabell 1) definierar vi det svenska akademiska skriftspråket som bestående av texter skrivna av och för akademiker, på svenska. Vi har således låtit det representeras av texter från disciplinerna humaniora och samhällsvetenskap, där den svenska akademiska produktionen är tillräckligt hög. Från dessa discipliner har vi sedan valt ämnesområden där förhållandevis många publikationer har funnits tillgängliga som fulltext-pdf:er genom SwePub. De humanistiska ämnesområdena är: *etnologi, filosofi, historia, konst, litteraturvetenskap, religion* och *språkve-*

4. Alla ordantal i tabellen är angivna i tusental.

tenskap, och de samhällsvetenskapliga: *ekonomi och näringsliv, juridik, medie- och kommunikationsvetenskap, psykologi, social och ekonomisk geografi, sociologi, statsvetenskap* och *utbildningsvetenskap*. Sammanlagt består korpusen av drygt 500 texter skrivna av fler än 450 olika författarkonstellationer.

3.2. Extraktion av akademiska ord

Till att börja med har allt korpusmaterial automatiskt annoterats med den teknik som används av Språkbanken, vilken bl.a. innefattar tokenisering, ordklasstaggning och lemmatisering (Borin et al. 2012b). Därtill har vi valt att utgå från lemmatiserade lexikala enheter i våra beräkningar; med *ord* menar vi alltså grundform inklusive samtliga böjningsformer. Vidare definierar vi *akademiska ord* som typiska för akademiska texter (**nyckelord**), där de är vanligt förekommande (**frekvens**) och jämnt spridda (**dispersion**) oberoende av ämnesområde (**utbredning**). Dessutom ingår de **inte i basordförrådet**.

För att automatiskt kunna extrahera dylika ord ur en akademisk korpus, måste definitionens alla kriterier operationaliseras, dvs. formuleras som regler vilka kan appliceras av ett datorprogram. [5]

5. Såväl vår definition av akademiska ord som de tekniska parametrarna i den automatiska extraktionsmodellen har bestämts genom att kombinationer av alla krav som tidigare använts för att extrahera akademiska ord testats och de resulterande listorna manuellt utvärderats. De slutliga kriterierna och tröskelvärdena är alltså de med vilka vi erhöll bäst precision (se vidare fotnot 6 och 7).

För att försäkra oss om att de akademiska orden inte ingår i det svenska basordförrådet har orden i SveAk filterats mot de 1000 mest frekventa orden i en korpus med lättlästa texter – *LäSBarT* (1,1 milj ord, Mühlenbock 2009). Några exempel på ord ur detta basordförråd är: *som, då, exempel* och *språk*.

För att objektivt skatta hur «vanligt» ett visst ord är i SveAk, använder vi ett frekvensmått som tar hänsyn till dispersion. Denna, s.k. *reducerade frekvens* (Savický & Hlaváčová 2002) ligger, enkelt uttryckt, närmare det absoluta frekvensvärdet om ordet är jämnt spritt i korpusen. Vidare räknar vi strikt med relativa frekvenser, för att kompensera för de olika ämneskorpusarnas varierande storlek.

De ord som uppvisar en reducerad frekvens på minst 15 förekomster[6] per miljon ord inom alla ämnesområden räknas som ämnesneutrala, och kvalificerar sig för den sista kontrollen. Denna urskiljer texttypiska nyckelord (Scott 1997) genom att ställa ordens (reducerade) frekvenser i SveAk mot motsvarande värden i en korpus med skönlitterära texter (2,5 milj. ord, Norstedtsromaner från 1999). Det akademiska nyckelordsförhållandet baserat på reducerade frekvenser, kallar vi för *akademiskt index*; ju högre detta värde är, desto mer akademiskt är ordet enligt vår definition. För att kandidera till den slutliga ordlistan måste det akademiska indexet uppgå till minst 1,1[7].

Slutligen har kandidatlistan rensats manuellt på oönskat brus, såsom förkortningar, textstrukturerande element (som *ii.*) och en del engelska ord, som taggaren inte lyckats identifiera som utländska.

6. Valet att sätta tröskelvärdet till 15 förekomster är heuristiskt baserat. 20 förekomster skulle resultera i ett alltför litet antal ord, medan 10 förekomster skulle riskera att öppna för alltför ovanliga eller ämnesspecifika ord.
7. Även detta tröskelvärde är grundat på heuristiska överväganden.

4. Presentation av ordlistan

Det arbete som beskrivs i avsnitt 3 ovan har resulterat i en samling med totalt 655 lexikala enheter. När dessa ordnas enligt fallande akademiskt index hamnar följande ord i topp:

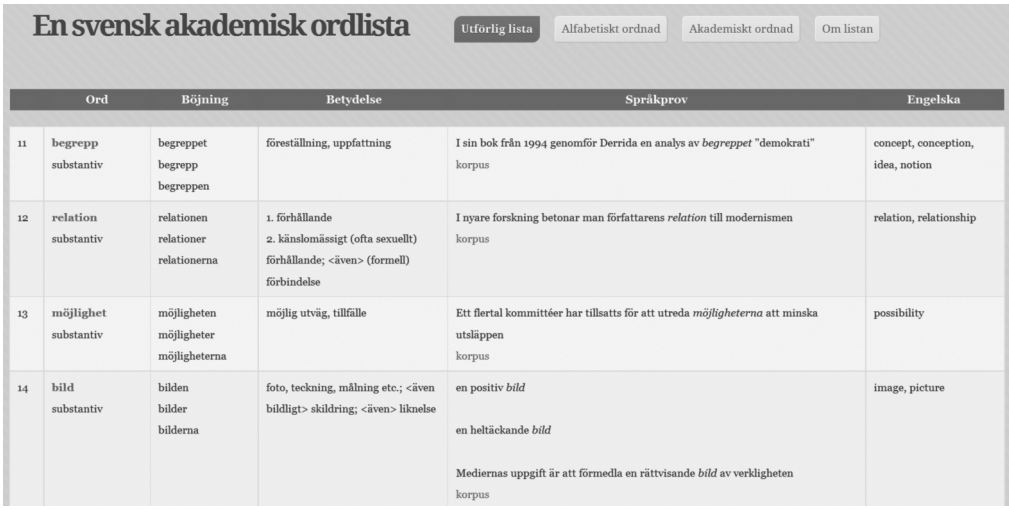
> *dock, studie, beskriva, social, enligt, innebära, samt, form, betydelse, fall*, *begrepp, relation, möjlighet, bild, utifrån skapa, analys, skillnad, utgöra, perspektiv*

Ordklassfördelningen i vår samling liknar i hög grad den i AKL (Paquot 2010). Huvuddelen (42 %) av listan består av substantiv. Vidare utgör verben 26 % och adjektiven 14 % av det totala antalet ord. Listan innehåller också många adverb, hela 8 %. (Jfr t.ex. den allmänspråkliga ordboken *Svensk ordbok utgiven av Svenska Akademien* 2009, som innehåller 68 % substantiv och 2 % adverb.)

Det faktum att vi i vårt arbete tagit fasta på lemman istället för ordfamiljer har tydliga konsekvenser för innehållet i listan. Formellt besläktade ord som *bedöma* och *bedömning*, *diskutera* och *diskussion* samt *omfatta*, *omfattande* och *omfattning* bildar egna uppslagsord. Genom vår metod synliggörs sålunda alla uppslagsord mer och – inte minst – de beskrivs på sina egna premisser. Vårt ställningstagande kan också kopplas till tanken att ordlistan i första hand ska användas vid produktion. Coxhead (2000) menar att bruket av ordfamiljer är befogat med tanke på att psykolingvistiska studier visat att morfologiska relationer mellan ord troligen finns lagrade i det mentala lexikonet. Paquot (2010) konstaterar att Coxheads resonemang håller för att presentera ordfamiljer för receptiva syften.

Däremot är presentationssättet föga meningsfullt vid produktion, då inte alla medlemmar i ordfamiljerna är lika användbara (jfr Gardner & Davies 2013:3f. som förespråkar användning av lemman i pedagogiska lexikala resurser).

Liksom SveAk är den svenska akademiska ordlistan fritt tillgänglig och nedladdningsbar via Språkbanken. Vidare är ordlistan införlivad i Språkbankens lexikala infrastruktur *Karp* (se vidare Borin et al. 2012a). Listans användargränssnitt framgår av figur 1.



Figur 1. Den akademiska ordlistans användargränssnitt.

I figur 1 återges ett visningsläge där uppslagorden är ordnade efter akademiskt index. Men användarna kan även välja att se orden i alfabetisk ordning.

De 100 översta uppslagorden bildar utgångspunkt för mer traditionella ordboksartiklar. Dessa ord är försedda med upp-

gifter om ordklass, böjning och betydelse, ett eller flera språkexempel samt engelska ekvivalenter. Exempelvis ges följande upplysningar om adverbet *dock*:

> **dock** (adverb) (oböjligt), 'i alla fall, ändå, likväl': *Efter ett par rosade novellsamlingar, som* dock *inte blev några försäljningssuccéer, började författaren att skriva romaner;* however, nevertheless, still, yet

Uppgifterna om ordklass, böjning och betydelse är hämtade från den nyligen uppdaterade *Lexins svenska lexikon* (2011, <http://lexin.nada.kth.se/lexin/>). Informationen har tillgängliggjorts via svenska Språkrådet som numera ansvarar för Lexinprojektet. I nuläget återges alla betydelser som anförs i Lexin, även om vissa betydelser torde vara vanligare än andra i akademiska texter. Ett exempel är verbet *uppfatta* som enligt ordboken kan betyda 'förstå, tolka' och 'lyckas höra'. Enligt vår bedömning är det främst den första betydelsen som är aktuell i SveAk. En systematisk granskning av vilken eller vilka betydelser som är vanligare i materialet hade givetvis bidragit till en bättre ordlista, men tyvärr saknades utrymme för ett sådant arbete inom projektets ramar.

Vidare är uppslagsorden försedda med ett eller flera redaktionella språkprov. Dessa är baserade på bruket i SveAk. Användarna kan också (via direktlänkar) klicka sig vidare från artiklarna till korpusen och på så sätt har de tillgång till fler – och autentiska – exempel utöver de enklare i artiklarna. Avslutningsvis är de engelska ekvivalenterna automatiskt hämtade från Lexins engelsk-svenska lexikon som tillhandahållits av Språkbanken.

5. Summering och framtida perspektiv

Den akademiska ordlista som presenteras här är tänkt som stöd för dem som behöver hjälp på vägen mot att erövra det svenska akademiska språket. I ordlistans förvalda presentationsform står orden i en ordning, där de mest typiska för akademiskt språkbruk står överst. Det innebär att listan kan rekommenderas till instudering i den ordning orden står.

De engelska översättningsekvivalenterna tillsammans med de svenska betydelseangivelserna gör att listan lämpar sig för såväl andraspråksinlärare som för modersmålstalande med liten erfarenhet av akademiskt språk. Länkningen till SveAk-korpusen ger vidare tillgång till en stor mängd autentiska exempel på hur varje ord kan användas.

Som redan antytts har projektets ekonomiska ramar inte tillåtit att alla uppslagsord försetts med utförligare information. Givetvis är det angeläget att åtgärda denna brist i framtiden. Därutöver kan vi se ett intresse för utbyggnad med återkommande akademiska fraser, i linje med vad som antytts i Carlund et al. (2012). Sett till ordlistans praktiska nytta, kan det finnas skäl att samarbeta med andra forskare vid framtagning av underlag för ordtester och undervisningsmaterial.

Användarstudier kan också bidra till kunskap om vidare utvecklingsbehov. Om det t.ex. skulle visa sig att information om ordfamiljer underlättar ordinlärningen, bör ordlistan byggas ut med sådana funktioner.

Litteratur

Bauer, L. & P. Nation (1993): Word families. I: *International Journal of Lexicography 6*, 253–279.

Borin, L., M. Forsberg, L.-J. Olsson & J. Uppström (2012a): The open lexical infrastructure of Språkbanken. I: *Proceedings of LREC 2012*. Istanbul: ELRA, 3598–3602.

Borin, L., M. Forsberg & J. Roxendal (2012b): Korp – the corpus infrastructure of Språkbanken. I: *Proceedings of LREC 2012*. Istanbul: ELRA, 474–478.

Carlund, C., H. Jansson, S. Johansson Kokkinakis, J. Ribeck, & J. Prentice (2012): An academic word list for Swedish – a support for language learners in higher education. I: *Proceedings of the SLTC 2012 workshop on NLP for CALL*. Linköping Electronic Conference Proceedings 80, 20–27.

Coxhead, A. (2000): A new academic word list. I: *TESOL Quarterly*, *34:2*, 213–238.

Coxhead, A. (2011): The academic word list 10 years on: Research and teaching implications. I: *TESOL Quarterly 45:2*, 355–362.

Gardner, D. & M. Davies (2013): A New Academic Vocabulary List. I: *Applied Linguistics 4*, 1–24.

Hyland, K. & P. Tse (2007): Is there an ”academic vocabulary”? I: *TESOL Quarterly 41:2*, 235–253.

Jansson, H., S. Johansson Kokkinakis, J. Ribeck & E. Sköldberg (2012): A Swedish academic word list: methods and data. I: R. V Fjeld & J. M. Torjusen (red.): *Proceedings of 15th EURALEX International Congress.* Oslo: University of Oslo, 955–960.

Johansson Kokkinakis, S., E. Sköldberg, B. Henriksen, K. Kinn

& J. Bondi Johannessen (2012): Developing Academic Word Lists for Swedish, Norwegian and Danish – a joint research project. I: R. V. Fjeld & J. M. Torjusen (red.): *Proceedings of 15th EURALEX International Congress.* Oslo: University of Oslo, 563–569.

Mühlenbock, K. (2009): Readable, legible or plain words − Presentation of an easy-to-read Swedish corpus. I: *Multilingualism, Proceedings of the 23rd Scandinavian Conference of Linguistics* (*Studia Linguistica Upsaliensia 8*). Uppsala: Acta Universitatis Upsaliensis, 325–327.

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

Paquot, M. (2010): *Academic vocabulary in learner writing. From extraction to analysis*. London & New York: Continuum.

Savický, P. & J. Hlaváčová (2002): Measure of word commonness. *Journal of Quantitative Linguistics 9*, 215−231.

Scott, M. (1997): PC analysis of key words – and key key words. *System 25/2*, 233−245.

Sköldberg, E. & S. Johansson Kokkinakis (2012): A och O om akademiska ord. Om framtagning av en svensk akademisk ordlista. I: B Eaker, L. Larsson & A. Mattisson (red.): *Nordiska studier i lexikografi 11.* Lund: Nordiska föreningen för lexikografi, 575–585.

Språkbanken. <http://spraakbanken.gu.se/>.

Svensk akademisk ordlista. < http://spraakbanken.gu.se/ao/>.

Svensk ordbok utgiven av Svenska Akademien (2009). Stockholm: Norstedts.

SwePub. <http://swepub.kb.se/>.

Wang Ming-Tzu, K. & P. Nation (2004): Word meaning in academic English: homography in the academic word list. I: *Applied Linguistics 25:3*, 291–314.

Judy Ribeck
doktorand i språkvetenskaplig databehandling
judy.ribeck@svenska.gu.se

Håkan Jansson
doktorand i nordiska språk
hakan.jansson@svenska.gu.se

Emma Sköldberg
universitetslektor, docent
emma.skoeldberg@svenska.gu.se

Inst. för svenska språket, Göteborgs universitet
Box 200, SE-405 30 Göteborg