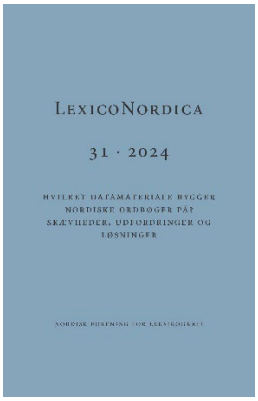
LexicoNordica

Titel:	Datatillgång, metodutveckling och lexikografiskt arbete vid Språkbanken Text
Forfatter:	Markus Forsberg & Louise Holmer
Kilde:	LexicoNordica 31, 2024, s. 61-79
URL:	https://tidsskrift.dk/lexn/issue/archive



Datatillgång, metodutveckling och lexikografiskt arbete vid Språkbanken Text

Markus Forsberg & Louise Holmer

This article discusses data access, methodology development, and lexicographical work at Språkbanken Text at the University of Gothenburg.

Although large, the different corpora accessible through Språkbanken Text's research infrastructure relevant for the work on contemporary dictionaries have mainly built upon newspaper texts from limited geographical areas, as well as texts from other genres, but often limited to specific time periods. However, since 2021, several joint efforts and cooperations have been initiated to develop and refine this aspect of Språkbanken Text's material. In this article, we describe the composition of the new corpus material, the development of new tools, and point out some possible research areas that have now appeared. One focus point is how the contemporary dictionaries SAOL (*Svenska Akademiens ordlista*) and SO (*Svensk ordbok utgiven av Svenska Akademien*) may benefit from new corpus material and new methods.

1. Inledning

I den här artikeln fokuserar vi på materialtillgång och metodutveckling vid Språkbanken Text vid Göteborgs universitet, framför allt i förhållande till arbetet med ordböckerna *Svenska Akademiens ordlista* (SAOL) och *Svensk ordbok utgiven av Svenska Akademien* (SO). Syftet är att ge en översikt över de nytillkomna material som finns tillgängliga via Språkbanken Text, särskilt genom ordforskningsplattformen Korp (Borin et al. 2012) i kombination med nyutvecklade verktyg för ordforskning. I artikeln diskuteras också möjliga framtida studier och lexikografisk utveckling.

För att närmare kunna beskriva de nutida materialen ges först

en introduktion till Språkbanken Text liksom det lexikografiska arbetet med SAOL och SO. Artikeln är i övrigt disponerad så att materialtillgången beskrivs och exemplifieras i huvudsakligen kronologisk ordning, och därefter fokuseras på metodutvecklingen. Avslutningsvis sammanfattas innehållet och vi ger förslag på framtida utvecklingsmöjligheter.

2. Språkbanken Texts lexikografiska verksamhet

Språkbanken Text är en del av en nationell forskningsinfrastruktur för språkliga data med syftet att stödja främst språkteknologisk och språkvetenskaplig forskning, men även annan forskning där språkliga data har en framträdande roll (Språkbanken Text 2024a). Språkbanken Text är placerad vid Institutionen för svenska, flerspråkighet och språkteknologi vid Göteborgs universitet. Sedan 2021 ingår dessutom den verksamhet och de lexikografer som arbetar med samtidsordböckerna SAOL och SO i Språkbanken Text (Borin & Holmer 2024). Den lexikografiska verksamheten vid Göteborgs universitet bedrivs alltså inom den mer renodlat språkteknologiska ramen såväl som inom det ordboks- och produktorienterade området, med mänskliga användare i åtanke.

2.1. Språkbanken Texts forskning

Språkbanken Text bedriver språkteknologisk forskning som ofta bidrar till att utveckla dess forskningsinfrastruktur. De språkteknologiska forskningsprojekten med lexikografiskt fokus har varit särskilt fruktsamma, speciellt ansatsen som gick under benämningen Svenskt Frasnät++ (Dannélls et al. 2021, jfr Borin & Holmer 2024:46) eftersom den knöt samman många tidigare lexikografiska projekt vid Göteborgs universitet. Dessa hade i några fall löpt över flera decennier, och i och med Svenskt Frasnät++ väcktes

därmed även tidigare arbeten, som var nära att falla i glömska, till liv.

De språkteknologiska arbetena vid Språkbanken Text med lexikografiskt fokus har länge haft ett särskilt uttalat syfte: att med hjälp av datorns hjälp göra storskalig automatisk ordanalys av stora mängder text, för att därmed göra texterna mer tillgängliga för forskning. Det gynnar också tillämpningen och det vetenskapligt grundade arbetet inom lexikografin.

Blickar vi bakåt ännu mer ser vi att Språkbanken Text är sprungen ur just ordboksverksamhet. Att den lexikografiska verksamheten vid institutionen sedan 2021 ingår i Språkbanken Text är på så vis en välkommen återgång till en tidigare ordning.

Att utveckla ordböcker med hjälp av språkteknologiska verktyg och språkteknologiskt förädlade textmaterial vid Göteborgs universitet går tillbaka ända till 1960-talet (Malmgren & Sköldberg 2013:125). Det arbetet fick sedermera en organisatorisk form 1975, när regeringen inrättade Logoteket (senare Språkbanken, nu Språkbanken Text) under Sture Alléns ledning. Mycket har hänt sedan dess, men den starka kopplingen mellan lexikografi och språkteknologi finns fortsatt kvar.

2.2. SAOL- och SO-redaktionen

SAOL och SO har länge haft sin redaktionella hemvist vid Göteborgs universitet. SAOL:s bakomliggande material flyttades till Göteborg från Lund på 1980-talet medan SO har sin grund i arbetet med Lexikalisk databas, till vilken grunden lades under 1960-talet (jfr Malmgren & Sköldberg 2013, se också Borin & Holmer 2024).

Traditionellt har det lexikografiska arbetet rent praktiskt gått till så att det redaktionella arbetet med en upplaga av ett visst verk har pågått under några år, sedan har den tryckta ordboken givits ut, och därefter har insatserna inriktats på nästa ordbok eller nästa

upplaga av SAOL eller SO. Arbetet med de två ordböckerna har alltså ofta gått omlott. Det här har också varit ett naturligt arbetssätt när det primära målet för verksamheten har varit att ge ut ordböcker i form av tryckta böcker.

När nu redaktionen för SAOL och SO ingår i Språkbanken Text har verksamheten blivit betydligt mer dynamisk. Detta sammanfaller också med att marknaden för tryckta ordböcker har gått tillbaka kraftigt. Ett exempel på det är att SO (2021) enbart publicerades digitalt (jfr Sköldberg 2022) och att den delvis historiska *Svenska Akademiens ordbok* (SAOB) enbart kommer att uppdateras digitalt. Planen för SAOL är dock att publicera den kommande upplagan som tryckt bok, liksom i olika elektroniska varianter. Det grundar sig bl.a. i SAOL:s långa tradition med tryckta ordböcker, liksom – faktiskt – efterfrågan från allmänheten.

De som utgör den aktiva forskargruppen involverade i SAOL och SO under perioden 2024–2028 består idag av runt tio lexikografer och språkteknologer, alla med olika specialkompetenser inom sina respektive områden.

3. Materialtillgång för svensk lexikografi

När ordboken *Svensk ordbok* 1986 (SOB) utvecklades vid Göteborgs universitet var det den första svenska ordbok som hade utarbetats med korpusbaserade metoder (Malmgren & Sköldberg 2013, Borin & Holmer 2024). Sedan dess har det lexikografiska arbetet vid Göteborgs universitet vilat på just digitala material och metoder, vilket numera är förhållandevis vedertaget i fråga om vetenskapligt grundade ordböcker över allmänspråket (jfr Atkins & Rundell 2008).

Det korpusmaterial som ordböckerna SAOL och SO (liksom deras föregångare) bygger på, har förändrats och utvecklats under åren. I detta avsnitt ges en översikt över dessa materials mest framträdande drag.

3.1. Press 65 – den första tidningskorpusen

Det tidigaste materialet som låg till grund för det stora forskningsprojektet *Nusvensk frekvensordbok* (NFO, Allén et al. 1970–1980) utgjordes av det som kom att kallas Press 65, fortfarande tillgängligt i Korp. Press 65 består av en miljon löpord från fem större svenska dagstidningar: Göteborgs Handels- och Sjöfartstidning, Svenska Dagbladet, Stockholmstidningen, Dagens Nyheter och Sydsvenska Dagbladet – Snällposten. Det insamlade materialet kom från noggrant utvalda delar av tidningarna, kategoriserade som utrikeskorrespondenters rapporter, kulturartiklar jämte recensioner och allmänna reportage. Ett antal texter som t.ex. sportartiklar, anonyma insändare, annonser med mera uteslöts medvetet under materialinsamlingen. Ytterligare en faktor som påverkade var att materialet skulle vara tillgängligt på hålkort för maskinell bearbetning (Språkbanken Text 2024b).

Press 65-innehållet består alltså av tidningstext från fem större morgontidningar där materialet är noggrant utvalt av den dåvarande forskargruppen i enlighet med ett antal kriterier. Denna korpus låg till grund för omfattande frekvensbaserade undersökningar om svenskans skriftliga ordförråd. De uppgifter som presenteras i *Tiotusen i topp* (Allén 1972) med frekvensinformation om svenskt skriftspråk kommer också från Press 65. Alltjämt finns Press 65 liksom dess efterföljare sökbara i Korp, något som bland annat tillåter diakroniska studier över det svenska ordförrådets utveckling.

3.2. SAOL och SO: äldre och nyare material

Från materialet i Press 65 har korpusunderlaget successivt utökats. För det lexikografiska arbetet med ordböckerna SAOL och SO liksom dess föregångare, har det framför allt varit aktuellt att använda textmaterial i form av korpusar med nyhetstexter. Efter Press

65 utökades materialet med ytterligare liknande nyhetskorpusar (t.ex. Press 76 på ca 1,3 miljoner löpord och Press 98 på ca 10,7 miljoner löpord) och därefter flera årgångar av morgontidningen Göteborgs-Posten, den sista årgången från 2013 med 16,9 miljoner löpord. I takt med att dessa material har tillgängliggjorts för forskning har de också legat till grund för *Nationalencyklopedins ordbok* (NEO, 1995–1996), SAOL 12 (1998), SAOL 13 (2006), SO (2009) och SAOL 14 (2015).

Allt eftersom korpusmaterialen har inkorporerats i Språkbanken Texts samlingar har också de material som det praktiska lexikografiska arbetet vilar på kunnat utökas sett till mängden, och det har även kunnat breddas genremässigt. I korpussamlingarna ingår också romankorpusar som består av både originalsvensk text och översättningar från framför allt engelska men även andra språk. Romanmaterialet är dock inte lika omfattande som tidningstextmaterialet, och även om det är välstrukturerat och t.ex. tillåter studier av enskilda författares språk, har materialet varit begränsat till att omfatta romaner från 1970–1990-talet. De ovannämnda romankorpusarnas storlek uppgår sammanlagt till ca 18,5 miljoner löpord. Till dessa kommer också hela textmängden från Litteraturbanken som också finns tillgänglig via Språkbanken Text, men den har inte tidigare använts aktivt i arbetet med ordböckerna. Förutom de moderna romankorpusarna och material från Litteraturbanken har det dessutom tillkommit ytterligare en romankorpus genom SAOB-redaktionens försorg (se avsnitt 3.3).

Under sent 1990-tal började texter publicerade på internet bli vanliga som skriftspråksform i det svenska samhället och det avspeglas även i Språkbanken Texts samlingar. De innehåller t.ex. ett antal omfattande korpusar med bl.a. bloggtexter. De 69 korpusarna med text från sociala medier innehåller sammanlagt ca 11,8 miljarder löpord, där ca 9,06 miljarder kommer från diskussionsforum och resten från t.ex. bloggar och twittertexter.

SAOL är en mer normativ ordlista medan SO är en mer

deskriptiv ordbok. Detta faktum har också inverkat på vilka texter som har legat till grund för materialvalen för de två ordböckerna. De något ledigare utformade bloggtexterna har främst använts i samband med SO-arbetet medan SAOL företrädesvis har grundats på det ofta mer vårdade skriftspråket från nyhetstexter. Båda ordböckerna har dock framför allt haft skriftspråk från nyhetstexter från de större svenska rikstäckande dagstidningarna, och i viss mån romantexter, som sin främsta materialkälla. För en översikt över arbetet med SO 2021, se Sköldberg (2022), och se Holmer (2022) för en något fylligare bild av framför allt utvalda tidnings-, roman- och bloggkorpusar, liksom hur dessa korpusar har använts i lexikala studier.

3.3 Språkbanken Texts nutida material

Språkbanken Text gör språkteknologiska bearbetningar av de material som läggs till i dess samlingar. Dessa bearbetningar syftar till att maximera tillgängligheten och nyttan för materialen som forskningsdata, utan att kränka lagar såsom upphovsrätten (jfr Bouma et al. 2024). Det kan till exempel handla om att skapa s.k. meningsmängder, där meningarna i ett material kastas om slumpvis, för att texterna inte ska vara intakta, samtidigt som referenserna till originalet bibehålls. Denna praktik liknar det som redan görs på internet av sökmotorer, där man kan få se ett utdrag av texten i sököversikten, medan man behöver besöka sidan för att få läsa hela texten.

Däri ligger även begränsningen i de material som finns tillgängliga via Språkbanken Text: det finns en tydlig slagsida mot material som är publicerade på internet, eftersom det är dessa material som verksamheten har teknisk tillgång till.

Men till skillnad från Språkbanken Text, har Kungliga biblioteket (KB) full tillgång till nästan allt som ges ut i Sverige, via lagarna om pliktexemplar (Kungliga Biblioteket 2024). Samtidigt har

KB inte laglig rätt att dela sina data vidare till Språkbanken Text – eventuella bearbetningar som måste göras av juridiska skäl för att datamängder ska kunna spridas fritt, behöver göras inom KB:s servermiljö. År 2021 påbörjades ett samarbete mellan Språkbanken Text och KB-labb, en forskningsinfrastruktursenhet vid KB, där just detta görs, något som har resulterat i nya, fritt tillgängliga, orddatamängder via Språkbanken Text, exempelvis i Korp: *Kubord*. Datamängderna fokuserar på lexikal information och syftar till att stödja ordforskning.

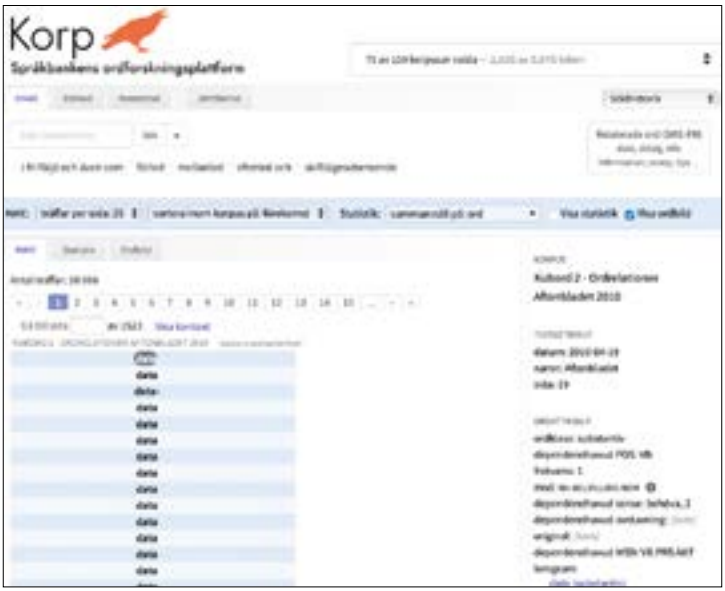
Kubord består av tidningstext från i huvudsak år 2010 till år 2021 och har hämtats från de större morgontidningarna Dagens Nyheter, Svenska Dagbladet och Göteborgs-Posten. Dessutom ingår Sydsvenska Dagbladet, som framför allt täcker Skåne, och Östgöta-Correspondenten, som ges ut i Östergötland (östra Mellansverige). Förutom dessa mer klassiska morgontidningar ingår också de två kvällstidningarna Aftonbladet (oberoende socialdemokratisk) och Expressen (obunden liberal). Till skillnad från de äldre materialen (Press 65 och liknande utvalda textmängder) innehåller Kubord-mängderna i princip de flesta delarna från de digitaliserade papperstidningarna, alltså både inrikes- och utrikesnyheter, sport, nöje m.m. De speglar därmed en större bredd av presstexterna än de tidigaste korpusarna.

I och med Kubord är det första gången som modernt kvällstidningsmaterial över huvud taget finns tillgängligt via Språkbanken Text. På sikt är det möjligt att det kommer att kunna gå att utöka antalet tidningar med olika typer av regional hemvist, vilket skulle balansera materialet ytterligare geografiskt. Det vanliga är annars att de största tidningarna anses tillräckliga, men de behandlar främst rikstäckande nyheter och har ofta storstäderna Stockholm, Göteborg och Malmö i fokus. Nyheter från t.ex. landsdelen Norrland kommer inte lika ofta med i tidningarna, och som en följd av det finns sådana nyhetstexter inte heller representerade i referenskorpusarna.

Delmängden Kubord 1 i Korp består av en fritt tillgänglig ordlistning med källreferenser som blivit språkteknologiskt förädlad för att innehålla sådant som ordklass, lemmatisering, ordbetydelser med mera. Kubord 1 är både nedladdningsbar via Språkbanken Texts datasida och tillgänglig via Korp.

Kubord 2 är en vidareutveckling av Kubord 1. Kubord 2 består av en listning med par av ord som är relaterade via en syntaktisk relation, exempelvis ett verbs direkta objekt eller substantivs framförställda attribut. Detta möjliggör att man exempelvis kan skapa en s.k. ordbild i Korp, vilken ger uppgifter om ordets syntaktiska relationer och liknande (jfr Borin et al. 2012:476). Sammanlagt bidrar Kubord med ca 6 miljarder ord.

För att illustrera visas i Figur 1 och 2 exempel från Kubord 2 i Korp via en sökning efter lemmat *data*.



Figur 1. Resultat vid sökning efter *data* i Kubord 2 när fliken KWIC är aktiverad (keyword in context)

I figur 1 visas sökresultatet för *data* med 38 066 träffar. Resultatet består alltså av sökträffar utan kontext, men med rik extra information (en delmängd visas till höger). Även om det kan se något fattigt ut fungerar det mesta av Korps funktionalitet för Kubord 2, förutom att användaren inte får någon kontext presenterad. I de fall användaren behöver gå till den ursprungliga texten finns källhänvisningar. För första träffen noteras Aftonbladet, 2010-04-19, sidan 19 (visas till höger i figur 1). Vi kan även observera t.ex. vilket syntaktiskt huvud som första träffen har, nämligen *behöva*.

Figur 2. Resultat vid sökning på *data* i Kubord 2 när fliken Ordbild är aktiverad

I figur 2 visas ordbilden för *data* i Kubord 2. Typiska attribut för *data* i Kubord-materialet är *extra*, *biometrisk* och *personlig*. Ordet *data* är också något som typiskt *köps*, *ingår*, *hämtas* och *samlas*.

Utöver Kubord-arbetet, kan tilläggas att redaktionen för SAOB i Lund nyligen utfört ett omfattande arbete med att skanna in romaner från perioden 1950 till 2007. Korpusen har fått namnet SAOB 1950 och finns i Språkbanken Texts samlingar, genom Korp och som öppet tillgänglig nedladdningsbar meningsmängd (Nationella språkbanken 2023). I och med den insatsen har mängden romanmaterial, ett tidigare ofta förbisett material, ökat avsevärt.

Dessutom täcker korpusen in även åren 1950 till 1980, decennier som i övrigt inte har varit särskilt välrepresenterade tidigare. Korpusen SAOB 1950 omfattar ca 50 miljoner ord, vilket innebär ett rejält tillskott till det befintliga romanmaterialet (se avsnitt 3.2).

3.4. Sammanfattning av materialutvecklingen

Både SAOL:s och SO:s nuvarande datamängder utgör resultatet av en mångårig tradition i kombination med nya samarbeten, nya material och nya metoder. Ett uppenbart problem ur materialsynvinkel är att tillgången till webbmaterial för forskningsändamål, som respekterar upphovsrätten och andra ekonomiska värden, minskar i takt med att betalväggar och andra tekniska lösningar begränsar tillgängligheten och därmed tillgången till data. Därför är det avgörande för svensk ordforskning att en verksamhet som Språkbanken Text ger sig in i strategiska partnerskap med de organisationer som har datatillgång via lagen eller på andra sätt har tillgång till relevanta data.

Från de tidigaste korpusarna med text från 1960- och 1970-talet omfattande någon miljon ord vardera, har tillgängliga material i Språkbanken alltså vuxit och omfattar nu många miljarder löpord fördelade på runt 300 korpusar med moderna material. Utöver det finns det numera också korpusar över fornsvenska, historiskt material från KB m.m. och diverse specialkorpusar som kan vara insamlade av enskilda forskare.

Förhoppningen är att på sikt kunna lägga till fler årgångar av moderna dags- och kvällstidningar i Korp, liksom tidningsmaterial från fler regioner så att materialet breddas geografiskt. Samtidigt får vi också konstatera att med moderna tidningskorpusar på 6 miljarder ord, som uppdateras löpande, i kombination med det tidigare och mycket omfattande textmaterialet, måste tillgången till svensk text i forskningssyfte sägas vara mycket god.

4. Nyutvecklade metoder

I samband med att de nya materialen har kunnat tillgängliggöras via Språkbanken Texts forskningsinfrastruktur (se avsnitt 3), har forskargruppen som arbetar med ordböckerna också kunnat utveckla särskilda verktyg för nyordsexcerpering och göra storskaliga jämförelser mellan befintliga ordboksmaterial och de nya korpusmaterialen. I följande avsnitt beskrivs först det nyordsverktyg som håller på att utvecklas internt inom gruppen och därefter arbetet med så kallade ordvektorer. Naturligtvis används också andra metoder och verktyg i kombination – förutom det här beskrivna nyordsverktyget används till exempel manuell excerpering av dagsaktuella texter, förslag på nya ord från allmänheten, loggfiler med icke-träffar från de elektroniska versionerna, jämförelser med motsvarande nordiska ordböckers listor över nytillkomna ord m.m. Här fokuseras dock på de nya metoder som har utvecklats sedan 2021.

4.1. Nyordsverktyg

Ett självklart led i att utveckla samtidsordböckerna är att uppdatera lemmalistan med nya ord, liksom att utmönstra föråldrade ord (jfr Diamond 2016). I grund och botten är maskinellt stöd för en sådan process enkel och ytterst beroende av relevant materialtillgång. Ordboksredaktionen vill undersöka vilka ord som ökar signifikant i frekvens i materialet över åren och vilka som minskar. Ord med ökad frekvens kan utgöra nyordskandidater medan ord med minskad frekvens kan utgöra strykningskandidater (jfr Berg, Holmer & Sköldberg 2010, Holmer et al. 2024).

Samtidigt blir det snabbt viktigt med en mer kritisk hållning till ordens frekvensinformation och deras ursprung. Om en signifikant frekvensökning kommer explosionsartat över ett par dagar

i bara en tidning, är det förmodligen en mycket sämre nyordskandidat än den vars frekvensökning är jämnare fördelad över tid och över flera material. Den senare fångar bättre vad forskarna och redaktionen är ute efter, nämligen att ordet har ökat i bruk i hela samhället.

Man kan också fråga sig vilken tidsenhet som är mest lämplig för frekvensjämförelser. Som regel väljs ofta år, inte minst för att korpusmaterial brukar delas upp årsvis, men det finns andra varianter, t.ex. ett glidande medelvärde. Därtill kommer hur det statistiskt kan avgöras att det alls skett en ökning och minskning. Alla varianter kommer med sina fördelar och nackdelar, så ett användbart nyordsverktyg behöver kunna hantera den variationen.

Sedan har vi frågan om analysenhet: vad menar vi egentligen med ett ord? Menar vi bara en ordform eller bör vi gruppera ordformer under ett visst lemma? Hur är det med de sammansättningar ordet ingår i, ska de också tas med i beräkningarna? Ska exempelvis *selfiemuseum* och *gymselfie* bidra till bedömningen av *selfie*?

Till sist har vi frågan hur en nyordkandidat ska presenteras. Ett idealiskt nyordsverktyg skulle gå hela vägen och också skapa ett ordboksartikelutkast med färdig information som böjningsinformation, typiska sammansättningar, språkexempel m.m.

Inom ordboksverksamheten jobbar vi med att utveckla ett nytt integrerat nyordsverktyg som tar sig an dessa frågeställningar och målsättningar. Hittills har detta verktyg framför allt kunnat användas för att vaska fram ett antal nyordskandidater. Följande handfull exempel, excerperade med hjälp av detta verktyg, utgör t.ex. potentiella nyord i SAOL och självständiga uppslagsord eller sammansättningsexempel i SO: *hållbarhetstänk*, *höjdrädsla*, *kärleksfront* (i frasen *på kärleksfronten*) *näringsjäst*, *samhällsutmaning*, *spontanköp*, *webbenkät* och *viltkamera*.

4.2. Ordvektorer

Under 2024 håller Språkbanken Text och KB-labb tillsammans på att utveckla en ny sorts Kubord-datamängd benämnd *Kubord-fasttext*, med ordvektorer. Enkelt uttryckt så kan man med ordvektorers hjälp få svar på frågan om vilka ord som har liknande språkliga kontexter i ett visst material.

Ordvektorer definieras algoritmiskt utifrån sin språkliga kontext på så vis att de ord som har liknande språkliga kontexter hamnar nära varandra matematiskt i en så kallad vektorrymd. För ett ord som *sjunga* kan det betyda att verb som *gnola, skråla* och *pratsjunga* ligger nära i vektorrymden, tillsammans med ord från andra ordklasser, som exempelvis *sång* och *skolkör*. Det senare är en indikation på en egenskap hos dessa ordvektorer, nämligen att syntaktisk position inte är en del av en ordvektors uppbyggnad. Så länge två ord förekommer i samma mening med liknande andra ord omkring sig hamnar de nära varandra.

I samarbetet använder vi oss av verktyget fastText (Bojanowski et al. 2017; verktygets namn skrivs med versalt T) för att skapa våra ordvektorer. Det som skiljer detta verktyg från andra verktyg som bygger ordvektorer, exempelvis Word2Vec, är att fastText konstruerar en ordvektor utifrån ordets delar. Det betyder i praktiken att ord med sammanfallande delar hamnar närmare varandra i vektorrymden, dvs. att ord som *sjunga* och *provsjunga* kommer närmare varandra på grund av att de delar *sjunga*.

Vad kan då dessa ordvektorer användas till i det lexikografiska arbetet? De kan användas för att identifiera lemmaluckor, hitta nya ordrelationer mellan lemman eller för att lokalisera typiska sammansättningar för ett ord, för att nämna några exempel. En ordvektor för *sommar* har exempelvis grannar som höst, vår och *vinter,* typiska sammansättningar som *sommarsäsong* och *försommar*, och även mer indirekta ordrelationer, som *båtsäsong*.

För mer detaljerade redogörelser om arbetet med ordvektorer

i en lexikografisk kontext, se Forsberg & Sköldberg (under utgivning).

5. Avslutande ord

I artikeln presenteras det arbete med nya material och nya metoder som har gjorts vid Språkbanken Text framför allt från och med år 2021. Fokus ligger på det lexikografiska arbete som ligger till grund för ordböckerna SAOL och SO. Vid Språkbanken Text pågår (eller har pågått) ytterligare projekt som är relaterade till arbetet med ordböcker och vars projektmedlemmar ibland är involverade i flera projekt samtidigt (se t.ex. Språkbanken Text 2024c).

Svenskan är trots allt ett välbeskrivet språk med flera ordböcker, grammatikor och korpusar. Som kontrast till artikelns fokus på svenskt skriftspråk kan nämnas arbetet med isländska talspråkskorpusar (Hilmisdóttir 2024) och arbetet med ett språk som meänkieli (Ahltorp et al. 2024). Att ständigt vilja utöka de svenska korpusmaterialen med ännu mer skriftspråk kan kanske tyckas onödigt. Vi ser dock möjligheterna i att, som nämnts, kunna öka den geografiska spridningen, något som möjliggör studier av t.ex. regional variation. Det finns också möjlighet till mer avancerade studier av ordförrådet över tid. Till de utökade materialen kommer en satsning på att metoder för identifiering och beskrivning av nyord.

Att ha tillgång till så dagsaktuella texter som möjligt är också helt nödvändigt för utvecklingen och uppdateringen av de två samtidsordböckerna SAOL och SO. Samarbetena med KB-labb har gjort att det lexikografiska arbetet kan baseras på ett mer kontinuerligt datainflöde av samtidstext än vad som har varit fallet tidigare, och i och med romankorpusen SAOB 1950 tillgängliggörs ett urval av 1900-talets skönlitterära texter i Korp.

Med tanke på att AI-genererade texter har blivit vanligare se-

dan åtminstone 2023, har det också blivit en viktig fråga att kunna skilja autentisk text från maskingenererad sådan, för att kunna göra en korrekt beskrivning av det svenska ordförrådet. Till viss del kan det avhjälpas genom tillgång till texter som man vet är (i stort sett) mänskligt producerade, exempelvis tidningstexterna tillgängliga via Språkbanken Text (t.ex Kubord 1 och Kubord 2). Men även i tidningstexter har användning av maskingenerad text börjat öka. En framtida utmaning blir att skilja dessa två texttyper åt, antingen via kunskap om materialet, till exempel genom att märka upp de delar som man vet att en tidning maskingenererar, eller genom att utveckla ny automatisk analys som försöker göra detsamma.

Avslutningsvis vill vi återigen lyfta hur central redaktionens inkorporering i Språkbanken Text har varit för båda verksamheterna, där det finns en växelvis draghjälp.

Litteratur

Ordböcker, korpusar och digitala resurser

Allén, Sture (1972): *Tiotusen i topp. Ordfrekvenser i tidningstext*. Stockholm: Almqvist & Wiksell.

KB-labb. <kb.se/samverkan-och-utveckling/kb-labb.html> (april 2024).

Korp = Språkbankens ordforskningsplattform. Version 9.0.6. <spraakbanken.gu.se/korp/> (augusti 2024).

Kungliga Biblioteket (2024). <kb.se/om-oss/det-har-gor-vi.html> (juli 2024).

Nationella språkbanken (2023). <sprakbanken.se/aktuellt/nyheter/2023-12-07-sprakteknologi-forenklar-arbetet-med-nya-saob> (juli 2024).

NEO = *Nationalencyklopedins ordbok*, band 1–3 (1995–1996). Höganäs: Bokförlaget Bra Böcker.

NFO = *Nusvensk frekvensordbok baserad på tidningstext* (1970–1980). Utarbetad av Sture Allén m.fl. Fyra band. Stockholm: Almqvist och Wiksell international (distr.).

SAOB = *Svenska Akademiens ordbok* (1898–2023). <www.saob.se/> (augusti 2024).

SAOL = *Svenska Akademiens ordlista över svenska språket*, 14 uppl. (2015). Stockholm: Norstedts.

SO = *Svensk ordbok utgiven av Svenska Akademien* (2021). <svenska.se/so/> (april 2024).

SOB 1986 = *Svensk ordbok* (1986). Göteborg: Språkdata & Esselte Studium AB.

Språkbanken Text 2024a. <spraakbanken.gu.se/> (mars 2024).

Språkbanken Text 2024b. <spraakbanken.gu.se/resurser/press65> (april 2024).

Språkbanken Text 2024c. <spraakbanken.gu.se/forskning> (augusti 2024).

Svenska.se = Svenska Akademiens ordboksportal. <svenska.se/> (mars 2024).

Annan litteratur

Ahltorp, Magnus, Lina Lejdebro Enwald, Elina Kangas, Jacob Larsson, Rickard Domeij & Gunnar Eriksson (2024): Språkteknologi för att samla in texter och analysera språket i korpusverktyg – hur gör man på meänkieli? I: *LexicoNordica* 31 (denna volym).

Atkins B.T. Sue & Michael Rundell (2008): *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.

Berg, Sture, Louise Holmer & Emma Sköldberg (2010): Time to say goodbye? On the exclusion of solid compounds from the Swedish Academy Glossary (SAOL). I: *Proceedings of the XIV Euralex International Congress (Leeuwarden, 6–10 July 2010)*. Leeuwarden: Fryske Akademy. 567–576.

Bojanowski, Piotr, Edouard Grave, Armand Joulin & Tomas Mikolov (2017): Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* (2017) 5, 135–146.

Borin, Lars, Markus Forsberg & Johan Roxendal (2012): Korp – the corpus infrastructure of Språkbanken. I: *Proceedings of LREC 2012. Istanbul: ELRA*. 474–478.

Borin, Lars & Louise Holmer (2024): Tradita innovare, innovata tradere. The Gothenburg way of computational lexicography. I: *Proceedings of the Huminfra Conference* (HiC 2024). 41–50.

Bouma, Gerlof, Markus Forsberg, Justyna Sikora & Emma Sköldberg (2024): Konsten att bedriva svensk ordforskning utan att kränka upphovsrätten. I: *Proceedings of the Huminfra Conference* (HiC 2024). 161–167.

Dannélls, Dana, Lars Borin & Karin Friberg Heppin (2021): *The Swedish FrameNet++ Harmonization, integration, method development and practical language technology applications*. Amsterdam: John Benjamins.

Diamond, Graeme (2016): Making Decisions about Inclusion and Exclusion. I: Philip Durkin (ed.): *The Oxford Handbook of Lexicography*. Oxford. 532–545.

Forsberg, Markus & Emma Sköldberg (Under utgivning). Ord med liknande kontext sökes! Om ordvektorers roll i svensk lexikografi.

Hilmisdóttir, Helga (2024). Talspråkskorpusar som resurs för isländska ordböcker. I: *LexicoNordica* 31 (denna volym).

Holmer, Louise (2022): Neutrala substantiv på *-ande* i text och ordbok. Meijerbergs arkiv 47. Göteborg.

Holmer, Louise, Ann Lillieström, Emma Sköldberg & Jonatan Uppström (2024): SAOL och svensk språkvetenskaplig infrastruktur – nu och i framtiden. I: *Proceedings of the Huminfra Conference* (HiC 2024). 68–75.

Malmgren, Sven-Göran & Emma Sköldberg (2013): The Lexicography of Swedish and other Scandinavian Languages. I: *International Journal of Lexicography*, 26(2), 117–134.

Sköldberg, Emma (2022): Andra upplagan av Svensk ordbok: förutsättningar och redaktionella val. I: *LexicoNordica* 29, 139–152.

Markus Forsberg
Föreståndare för Språkbanken Text
Inst. för svenska, flerspråkighet och språkteknologi
Göteborgs universitet
Box 200
SE-40530 Göteborg

Louise Holmer
Forskare, lektor
Inst. för svenska, flerspråkighet och språkteknologi
Göteborgs universitet
Box 200
SE-40530 Göteborg