Konsten att bedriva svensk ordforskning utan att kränka upphovsrätten

Gerlof Bouma[1], Markus Forsberg[1], Justyna Sikora[2] and Emma Sköldberg[1]

[1]*Institutionen för svenska, flerspråkighet och språkteknologi, Språkbanken Text, Göteborgs universitet*
[2]*KB-labb, Kungliga biblioteket*

Abstract

Vi beskriver KB-labb och Språkbanken Texts samarbete för att underlätta ordforskning på de upphovsrättsskyddade korpusar som finns i Kungliga bibliotekets samlingar. Satsningen har hittils lett till två öppna datasamlingar, Kubord 1 och 2, som ger tillgång till ordstatistik och ordsamförekomststatistik. Vi beskriver även Kubord-fastText, en samling vektormodeller som är baserade på samma korpusar, som är under utveckling.

Keywords

ordforskning, ordvektorer, ordstatistik, lexikografi, tidningstext

1. Inledning

Vid Göteborgs universitet bedrivs det sedan flera decennier tillbaka forskning kring ämnena lexikografi, lexikologi och fraseologi. Digitaliserade textsamlingar – korpusar – har länge spelat en avgörande roll inom denna forskning. Korpusarna är även, och har länge varit, centrala för den ordboksverksamhet som pågår inom Språkbanken Text och som bland annat mynnar ut i Svenska Akademiens samtidsordböcker (SAOL [1] och SO [2]; se vidare [3]). Underlaget för senare upplagor av de aktuella ordböckerna har i huvudsak bestått av redigerade texter i form av tidningstext, men även en del romaner. Av upphovsrättsliga skäl är dock tillgången till digitaliserade tidningar och romaner begränsad och det är givetvis ett bekymmer för de forskare som studerar svenskans ordförråd och för de språkvårdare som till exempel ska ge rekommendationer kring ordval. Ordforskningen behöver tillgång till stora korpusar, med olika slags texter, särskilt moderna texter men också texter från olika tidperioder, för att kunna undersöka ovanliga ord, ords spridning, nya ord och deras utveckling både i betydelse och i form, för att nämna några aspekter. Men bristen på korpusar drabbar inte bara språkforskningen utan också annan humanistisk och samhällsvetenskaplig forskning, även om det inom dessa inriktningar sällan är själva språket som är i fokus, utan det som språket förmedlar.

För att motverka denna situation pågår det nu ett samarbete mellan KB-Labb och Språkbanken Text, som syftar till att tillgängliggöra upphovsrättsskyddade texter – särskilt moderna pressmaterial – utan att hamna i konflikt med upphovsrätten, på sätt som främjar forskningen. Arbetet går också ut på att vidareutveckla språkteknologiska metoder som kan stärka den vetenskapliga

Huminfra Conference 2024, Gothenburg, 10–11 January 2024.
gerlof.bouma@gu.se (G. Bouma); markus.forsberg@svenska.gu.se (M. Forsberg); justyna.sikora@kb.se (J. Sikora); emma.skoldberg@svenska.gu.se (E. Sköldberg)

kvaliteten på ordforskning, och inte minst ordboksarbete. Samarbetet har bland annat resulterat i datasamlingarna Kubord 1 och Kubord 2 som, exempelvis via forskningsverktyget Korp, är fritt tillgängliga hos Språkbanken Text. Därtill är ännu en datasamling, Kubord-fastText, på väg att göras publik. I denna artikel kommer resultaten av det pågående samarbetet att presenteras och diskuteras. Vi kommer även att blicka framåt och resonera kring hur det pågående samarbetet kan fördjupas och stärkas ytterligare i framtiden.

2. Kubord 1, Kubord 2 och Kubord-fastText

2.1. Kubord 1

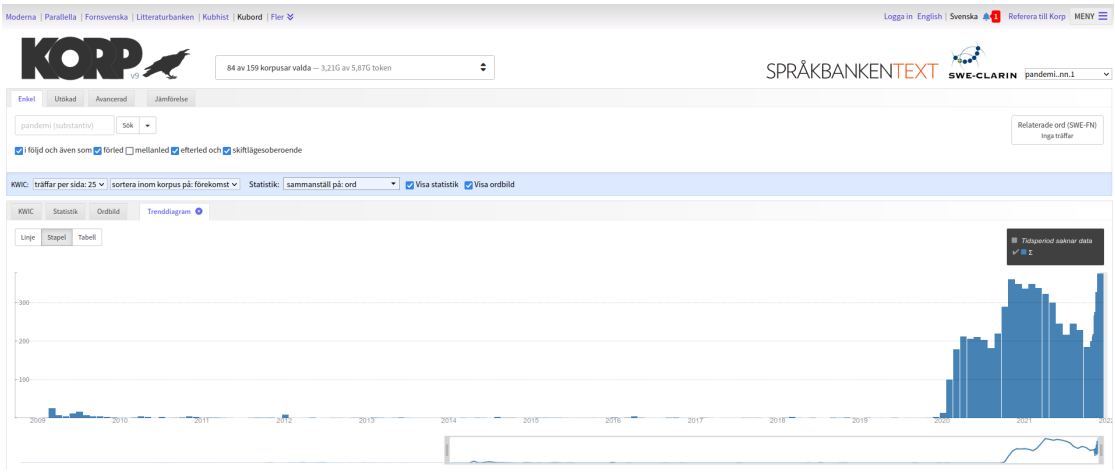
Det första konkreta resultatet av det pågående samarbetet går under beteckningen Kubord 1, en datasamling som finns fritt tillgänglig via Språkbanken Text.[1] Kubord 1 bygger på drygt 80 delsamlingar med olika pressmaterial, sammanlagt lite mer än 3 miljarder tokens. Tidningsmaterialen är publicerade mellan åren 2000–2021 i bland annat morgontidningar (*Dagens Nyheter* och *Svenska Dagbladet*), kvällstidningar (*Aftonbladet* och *Expressen*) och landsortspress (*Östgöta Correspondenten*). Att Kubord 1 enbart innehåller tidningsmaterial kan tyckas begränsande, men tidningar innehåller trots allt många olika slags texttyper. I pressmaterial behandlas också en rad olika ämnen vilket resulterar i en stor spridning (såväl ämnesmässig som stilistisk) bland de ord som påträffas i korpusarna.

På grund av upphovsrätten kan Kubord 1 inte göras sökbar på samma sätt som andra korpusar inom Språkbanken Texts forskningsinfrastruktur, via exempelvis Korp: det går att slå upp ord, men förekomsterna visas inte i vanligt konkordansformat. Det man kan få, däremot, är detaljerad frekvensinformation. Som allt annat textmaterial inom Språkbanken Text är Kubord-materialet försett med metadata som källhänvisningar och är automatiskt berikat med hjälp av Språkbankens analysplattform Sparv [4]. All denna information kan användas för att förfina sökningarna och statistiken. För att kunna analysera ett ord, till exempel bestämma ordets grundform, ordklass eller betydelse, krävs det för det mesta att man har tillgång till den språkliga kontexten. Med andra ord är vi beroende av kontexten i det automatiska analyssteget. Men i och med att KB inte har möjlighet att dela materialet med Språkbanken Text, på grund av upphovsrätten, körs Sparv därför på KBs servrar. När analysen är gjord, tas kontexten bort. Det går därmed inte att ens återskapa textfragment, och på så sätt kan vi tillgängliggöra en förädlad datasamling med ord utan att strida mot upphovsrätten.

För att illustrera vilken sorts information man kan få ut ur Kubord 1, återges resultaten av en sökning i figur 1, där vi tittar på ett ord vars användning har ökat lavinartat de senast åren, nämligen substantivet *pandemi*. Ordet har använts drygt 86 000 gånger i de aktuella texterna och det används framför allt i bestämd form singularis. En sökning i Kubord 1 visar bland annat också att den vanligaste sammansättningen med *pandemi* som för- och efterled är *pandemiåret*, *pandemilagen* och *pandemitider* respektive *viruspandemi* och *influensapandemi*. I figur 1 visas ett trenddiagram för ordet, inklusive användning i sammansättningar. Ökningen i förekomster syns mycket tydligt med start i 2020, det vill säga, när Covid-19-pandemin bröt ut.

I samband med exempelvis lexikografiskt arbete är det ofta viktigt med den språkliga kontexten.

[1] Kubord 1 finns tillgänglig via https://spraakbanken.gu.se/resurser/kubord.



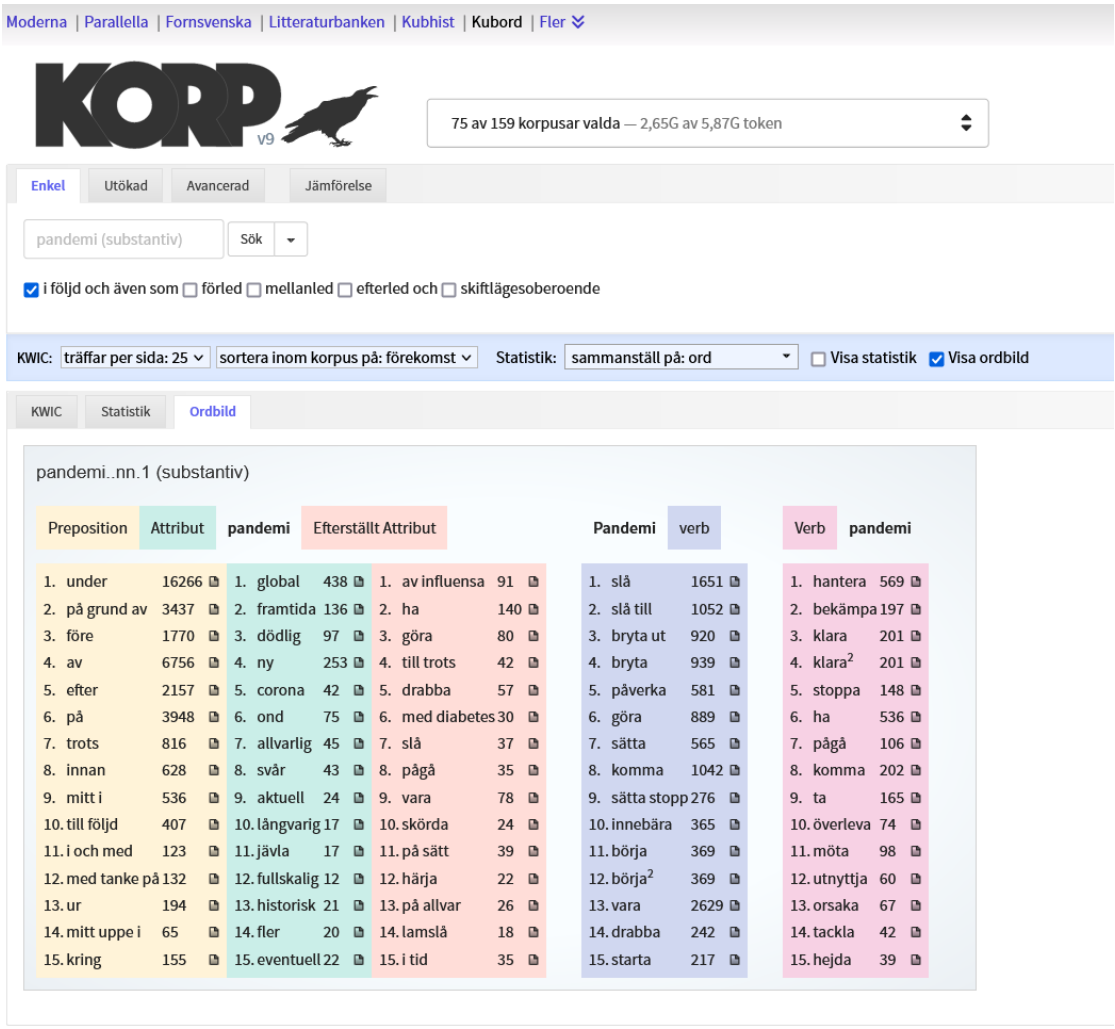
Figur 1: Trenddiagram för ordet *pandemi* i Kubord 1.

Det sammanhang som orden används i behövs för att man ska kunna avgöra ords betydelse(r) och fastslå i vilka konstruktioner orden typiskt uppträder i. Källhänvisning finns förvisso i Kubord 1, men det kräver en hel del extra arbete att finna rätt på den faktiska källan för en viss förekomst. Kubord 1:s utformning begränsar därmed givetvis användningsområdena inom ordforskningen. Samtidigt kan forskaren komma en bra bit på väg med frekvensangivelser gällande berikade ordformer. Exempelvis räcker Kubord 1 till att undersöka likheter och skillnader i ordförrådet mellan två årgångar av en och samma tidning. Sådana jämförelser är användbara inom bland annat det nyordsarbete som bedrivs inom ordboksprojektet (se vidare [5, s. 74]). Inom fältet mat och dryck är ord som *streetfood*, *charkbricka*, *gastropub*, *miso*, *pommes*, *prosecco*, *salsiccia*, *syrah* och *teriyaki(sås)* exempel på ord som kommit starkt under senare år. Många av dessa ord, som huvudsakligen är substantiv, säger något om hur svenska ord bildas. De säger även något om vår samtid, till exempel om våra matvanor och hur det omkringliggande samhället förändras över tid. Nya ord som dessa kan antingen bli uppslagsord i kommande upplagor av ordböcker eller fungera som språkexempel, då i form av sammansättningar eller avledningar i aktuella lexikografiska verk (se vidare [6]). Jämförelser i ordförråd går inte bara att göra mellan årgångar av tidningar i Kubord 1:s material, men också mellan material från Kubord och andra material i Korp, exempelvis tagna från sociala medier eller webbsidor.

Innehållet i Kubord 1 fungerar även som stöd vid utmönstring av uppslagsord. Just nu pågår ett arbete med att jämföra ordförrådet i de aktuella dagstidningarna med förteckningen med uppslagsord i SAOL för att de hur väl ordförrådet i ordlistan speglar de ord som faktiskt används i dagspressen av idag.

2.2. Kubord 2

Kubord 1 är, som sagt, begränsat till statistik över enstaka ord. I den andra samlingen som KB-labb och Språkbanken Text har tagit fram, Kubord 2, får man också tillgång till statistik över ordpar som står i syntaktisk relation till varandra, såsom verb-subjekt eller substantiv-attribut.



Figur 2: Ordet *pandemi* i Kubord 2.

Med andra berikas Kubord-orden därmed med viss kontextuell information. Även Kubord 2 är fritt tillgänglig via Språkbanken Text.[2] Samlingen är baserad på i princip samma tidningsmaterial som Kubord 1, och har försetts med samma metadata och språkteknologiska analys.

Informationen om grammatiskt relaterade ordpar möjliggör så kallade *ordbilder* för de ord som förekommer i korpusarna. Ordbildsvisningen är ett mycket användbart verktyg inte minst när ordforskaren vill undersöka en lexikal enhets typiska medspelare i en sats. Låt oss åter ta substantivet *pandemi* som exempel. I de pressmaterial som ingår i Kubord 2 används detta substantiv som sagt inte mindre än drygt 86 000 gånger. Med hjälp av ordbildsvisningen kan forskaren skapa sig en överblick över alla dessa fall och lättare se mönster i hur ordet brukar användas. Visningen effektiviserar och förbättrar därmed analysarbetet avsevärt (se också [5,

[2]Kubord 2 finns tillgänglig via https://spraakbanken.gu.se/resurser/kubord2

s. 76]).

Ordbilden för *pandemi* i Kubord 2 visas i figur 2. Av ordbilden i figuren framgår bland annat att ordet *pandemi* preciseras av adjektiv som *global*, *dödlig* och *ny*. Vidare står ordet som objekt till verb som *hantera*, *bekämpa*, *klara*, *stoppa* och *pågå*. Ordbilden visar också att ordet förekommer som subjekt till verb som *slå till* och *bryta ut*.

Ett av de aktuella verben som brukar uppträda tillsammans med *pandemi* är alltså partikelverbet *bryta ut*. En ordbildssökning i Kubord 2 på just det ordet tydliggör i sin tur att detta verb har flera betydelser (se vidare bland annat [5, s. 77] om värdet av ordbilder vid identifiering av ords olika betydelser). För det första kan *bryta ut* utgöra en synonym till 'lösgöra ur en större helhet'. Ett återkommande objekt till ordet är då substantivet *del*. För det andra kan verbet betyda 'inleda', 'starta'. Verbet uppträder då tillsammans med subjekt eller objekt som *brand*, *krig*, *slagsmål*, *protest*, *strejk* och *orolighet*. Som synes rör det sig ofta om negativt laddade ord. Ett objekt som avviker i ordbilden är *jubel*. Verbet *bryta ut* kan då sägas betyda 'brista ut'.

En uttömmande beskrivning av såväl *pandemi* som *bryta ut*, dess betydelser och kombinatoriska drag, ska sålunda ge information om bland annat detta. Uppgifter om hur de aktuella orden kombineras med andra ord är mycket viktiga uppgifter i ordböcker, i synnerhet sådana som vänder sig till inlärare av ett språk. I definitionsordboken SO har beskrivningen av olika slags ordkombinationer nått långt men den kan onekligen förbättras, inte minst genom att fler återkommande ordkombinationer läggs till (se vidare bland annat [7]). I samband med det arbetet kommer användningen av ordbildsvisningen i Kubord 2 att spela en viktig roll.

2.3. Kubord-fastText

Hittills har vi uppehållit oss kring Kubord 1 och 2 som redan är öppet tillgängliga via Språkbanken Text. Detta avsnitt handlar om Kubord-fastText, som är en resurs under utveckling.

Inom projektet *Svenska Akademiens samtidsordböcker* har vi tidigare genomfört pilotstudier med ordvektorer, som bland annat kan förse lexikografer med ord vars språkliga kontexter liknar varandra, och utforskat hur användningen av sådana kan berika det lexikografiska arbetet. Av [8] framgår metodens användbarhet i lexikografiska sammanhang på flera sätt. Ordvektorerna kompletterar de uppgifter lexikograferna får fram med hjälp av konkordanser och ordbilder. Bland grannarna i de vektorrymder som granskas återfinns till exempel ofta semantiskt besläktade ord till det ord som studeras. Dessa grannord kan ligga till grund för tillägg av fler hänvisningar till synonyma, antonyma och kohyponyma ord inom ordboksartiklarna. Metoden kan därmed, på ett förhållandevis objektivt och datadrivet sätt, förtydliga kopplingar mellan befintliga uppslagsord i ordboken och sådana som läggs till i samband med en revidering. Bland grannarna finns det också många sammansättningar som kan tjäna som morfologiska språkprov i ordboksartiklar. Därtill kan ordvektorer ringa in olika slags semantiska fält och ge information om ords värdeladdning.

För att konkretisera återvänder vi en sista gång till substantivet *pandemi*. Bland grannarna i vektorrymden till detta ord finner man ord som *coronapandemi*, *covidpandemi*, *pandemiår*, *pandemivåg* och *pandemiläge* (som alla innehåller det aktuella ordet), men också andra ord som är något mer avlägsna (till form och/eller innehåll) men ändå klart relaterade till det aktuella substantivet och den samhälleliga krissituation som pandemin orsakade. Exempel är *epidemi*, *hälsokris*, *folksjukdom*, *covidrestriktioner*, *lockdown*, *smittpridning*, *smittovåg*, *platsbrist* och *vaccinbrist*.

Ett problem i samband med dessa pilotstudier har dock varit en bristande kontroll över vad som utgör ett ord i vektorrymden, vilket begränsar användningen och därtill ger en hel del brus i vektorrymder. Vidare är det önskvärt att ha ökad kontroll över vilka korpusmaterial som ordvektorerna baseras på. Inte bara för att det är centralt att ha tillgång till källhänvisning, utan även för att kunna jämföra ordvektorer över exempelvis tid och material. Att försöka ta sig an dessa problem har varit centralt i arbetet med Kubord-fastText. Ett viktigt mål för vår pågående utredning är att få fram en väl underbyggd rekommendation för hur vektorrymderna ska vara beskaffade för att vara så användbara som möjligt.

Som namnet redan anger använder vi oss av fastText [9] för att skapa våra vektorrymder, vilket är en metod som skapar vektorrymder där varje ord representeras utifrån dess delar. Detta står i kontrast till metoder som alltid behandlar ord som en helhet, se till exempel [10] som använder sig av en sådan metod i ett lexikografiskt sammanhang. Användningen av fastText möjliggör hantering av ord som inte har observerats i träningsdatan, så länge delar av ordet har blivit det, vilket är en viktig egenskap för svenska språket med sin rika produktion av sammansättningar.

3. Sammanfattning och framåtblick

I den här artikeln har vi i korthet redogjort för resultaten av ett pågående samarbete mellan KB-labb och Språkbanken Text. Arbetet går ut på att, inom ramarna för upphovsrättsskyddet för olika källmaterial, utveckla nya datasamlingar som är av största möjliga nytta för svensk ordforskning.

De samlingar som har tagits fram har begränsningar, bland annat i och med att ordens kontexter inte visas upp, men samlingarna utgör trots det ett viktigt bidrag till ordforskare. Inte minst har arbetet kring dessa samlingar höjt den vetenskapliga kvaliteten på det nyordsarbete som bedrivs inom projektet Svenska Akademiens samtidsordböcker vid Göteborgs universitet. Samarbetet mellan KB-labb och Språkbanken Text banar också väg för metodutveckling på mer generell nivå. Exempelvis har det redan resulterat i studier av ordvektorers roll inom lexikografin.

I nuläget sträcker sig Kubord-materialen fram till och med år 2021, men materialen kommer att kompletteras med nyare pressmaterial allt eftersom, vilket är viktigt för exempelvis ordforskare med fokus på de ord som tillkommit eller vunnit terräng på senare år och hur dessa används.

4. Efterord

Detta arbete har möjliggjorts av *Nationella språkbanken* och *HUMINFRA*, båda finansierade av Vetenskapsrådet (20182024, kontrakt 2017-00626; 20222024, kontrakt 2021-00176) och deras samarbetsorganisationer samt av projektet *Svenska Akademiens samtidsordböcker*, finansierat av Svenska Akademien.

Referenser

[1] Svenska Akademiens ordlista, 14 ed., 2015. Tillgänglig via https://svenska.se.

[2] Svensk ordbok utgiven av Svenska Akademien, 2 ed., 2021. Tillgänglig via https://svenska.se.

[3] S.-G. Malmgren, E. Sköldberg, The lexicography of Swedish and other Scandinavian languages, International Journal of Lexicography 26 (2013) 117–134.

[4] M. Hammarstedt, A. Schumacher, L. Borin, M. Forsberg, Sparv 5 User Manual, Technical Report, Göteborg, 2022.

[5] A. Kilgarriff, Using corpora as data sources for dictionaries, in: H. Jackson (Ed.), The Bloomsbury Handbook of Lexicography, Bloomsbury Academic, London, 2013, pp. 71–88.

[6] E. Sköldberg, Hur fångar vi upp svenskans nya ord med hjälp av kubord, Språkbanksbloggen, 2022. Tillgänglig via https://spraakbanken.gu.se/blogg/20221128-hur-fangar-vi-upp-svenskans-nya-ord-med-hjalp-av-kubord.

[7] E. Sköldberg, Phraseological theory, evidence in corpora and lexicographical practice: on collocations in a monolingual dictionary of Swedish, in: K. Blensenius (Ed.), Valency and constructions. Perspectives on combining words, number 46 in Meijerbergs arkiv för svensk ordforskning, Meijerbergs institut, 2022, pp. 155–182.

[8] M. Forsberg, E. Sköldberg, Ordvektorer i lexikografiskt arbete, in: E. Volodina, D. Dannélls, A. Berdicevskis, M. Forsberg, S. Virk (Eds.), Live and Learn. Festschrift in honor of Lars Borin, Institutionen för svenska, flerspråkighet och språkteknologi, Göteborg, 2022, pp. 37–41.

[9] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, Transactions of the Association for Computational Linguistics 5 (2017) 135–146. URL: https://aclanthology.org/Q17-1010. doi:10.1162/tacl_a_00051.

[10] N. H. Sørensen, S. Nimb, Word2Dict – Lemma Selection and Dictionary Editing Assisted by Word Embeddings, in: J. Čibej, V. Gorjanc, I. Kosem, S. Krek (Eds.), Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts, Ljubljana University Press, Faculty of Arts, Ljubljana, Slovenia, 2018, pp. 819–826.