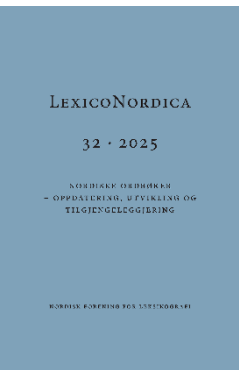


LexicoNordica

Titel:	Framtidssäkrade samhällsord i Lexin	
Forfatter:	Magnus Ahltop & Laila Höglund	
Kilde:	LexicoNordica 32, 2025, s. 17-36	
URL:	https://tidsskrift.dk/lexn/issue/archive	

© 2025 LexicoNordica og forfatterne

Betingelser for brug af denne artikel

Denne artikel er omfattet af ophavsretsloven, og der må citeres fra den. Følgende betingelser skal dog være opfyldt:

- Citatet skal være i overensstemmelse med „god skik“
- Der må kun citeres „i det omfang, som betinges af formålet“
- Ophavsmanden til teksten skal krediteres, og kilden skal angives, jf. ovenstående bibliografiske oplysninger.

Framtidssäkrade samhällsord i Lexin

Magnus Ahlthorp & Laila Höglund

Much has happened since the Lexin dictionary series for immigrants learning Swedish was first launched in Sweden in the early 1980s. Technical innovations, increased immigration from non-European countries, together with an ongoing digitisation have led to increased demand for advanced Swedish literacy skills and more rapid integration. An increased number of students going through university education today, along with a growing importance of digital literacy across the labour market in previously mainly manual industrial roles, now means an increased need for more specialised terminology for immigrants learning Swedish.

As language usage changes to reflect societal change, regular revision is important to keep contemporary dictionaries up-to-date. Most affected are the so-called *samhällsord* ‘societal terms/words’ in Lexin, words covering specific issues, celebrations of cultural importance, and names of institutions in Swedish society. These are words with additional factual information, which have been included in the Lexin dictionaries as they are “of importance to students wishing to familiarise themselves with modern Swedish society”, as stated in the preface of the Lexin Swedish-English dictionary edition from 1997. “As society evolves, however,” it concludes, “these explanations will gradually become outmoded”. This article focuses on ways of identifying and updating this particular group of lemmas through categorisation and partial automatisisation.

1. Inledning

Lexikonserien Lexin tillkom inom ett svenskt forsknings- och utvecklingsprojekt som initierades 1977 med syfte att utreda och skapa förutsättningar för produktion av språklexikon för invandrare och flyktingar som studerar svenska. Svenskkunskaper är viktiga för invandrares möjligheter att integreras i samhället och språk-

lexikon är ett betydelsefullt hjälpmedel när man ska lära sig ett nytt språk. (Se bl.a. SOU 1983:58, Beijer & Fre Woldu 1997:4ff. För utförligare bakgrund och lägesbeskrivning, se Gellerstam 1999:3.) Språklexikon var även ett viktigt hjälpmedel i genomförandet av riksdagsbeslutet om att främja aktiv tvåspråkighet.

Lexin förvaltas sedan 2007 av Språkrådet, Institutet för språk och folkminnen. I dag består Lexin av tryckta och digitala inlärarlexikon på ett tjugotal olika minoritetsspråk, en bildordbok samt tematiska videofilmer. Den primära målgruppen för Lexinlexikonen är nyanlända invandrare och flyktingar som lär sig svenska och lexikonen i serien är unika i att de i hög grad är utarbetade utifrån denna målgrupps behov och förutsättningar. Urvalet av uppslagsord är anpassat till den vardag de möter i Sverige, med arbete eller studier, familj, skola och bostad och utgörs av högrekventa och vardagliga ord ur allmänspråket, men även specifika ord som invandrare behöver känna till för att kunna orientera sig i det svenska samhället, ord som är vanligt förekommande i politiska eller samhälleliga sammanhang, exempelvis *skatt*, *fackförbund*, *riksdagsval* och *dagis*. Till dessa s.k. *samhällsord* finns förutom språklig information även sakupplysningar. (Se bl.a. Gellerstam 1999:5, Hult 2017:35.) Figur 1 visar artikeln för samhällsordet *medborgarskap* med en förklarande sakupplysning i webbgöransnittet lexin.se.

1.1. Bakgrund till projektet

Sedan Lexinseriens tillkomst i början av 1980-talet har världen förändrats och med den ordboksanvändarnas behov. Dagens digitala informationssamhälle ställer krav på avancerad litteracitet. Genom ny teknik, fri rörlighet och utbyte av varor och tjänster samt alltmer globala diskussioner och trender i sociala medier förändras språkbruket i snabb takt. En större medvetenhet kring social rättvisa i dagens Sverige påverkar också språket, och ord-



Figur 1: Samhällsorden i Lexin är försedda med extra sakupplysningar eller förklaringar.

boksartiklar där exempelvis föräldrade eller missvisande namn på grupper förekommer kan behöva revideras och uppdateras utan dröjsmål. Ord som söker beskriva samhället, dess beståndsdelar och kulturyttringar påverkas i högre grad av dessa språkförändringar än vardagliga substantiv som rör familj, hem eller husdjur.

Det är därtill av särskild vikt att kategorin av lemman som kallas *samhällsord* är aktuell och relevant för att underlätta för andraspråksinlärare att orientera sig i det svenska samhället. Andraspråksinlärare har ett specifikt lexikografiskt behov av att få förklaringar till en rad kulturspecifika förhållanden som inte självklart förstås av en utomstående och inkluderar ett urval av ord som rör relevanta lagar, samhällsinstitutioner och kulturfenomen (Tarp 1999:107, 2008:143). Samhällsorden är särskilt relevanta att lära sig tidigt för att lära känna och bli en del av det nya samhället (Loenheim & Hult 2020:14). Samhällsinstitutioner, kulturföretelser och lagar i urvalet bör därför vara aktuella och relevanta för Lexinanvändarna.

Åtskilligt har gjorts för att bibehålla Lexinlexikonens relevans, aktualitet och användaranpassning, senast genom den uppdatering och genomgripande metalexikografiska förbättring av den svenska orddatabasen som genomfördes vid Göteborgs universitet 2008–2010 (Malmgren 2012). Även samhällsorden har reviderats genom åren, och urvalet har kompletterats framför allt inom det ekonomiska området (Beijer & Fre Woldu 1997:9), men någon genomgripande översyn och revidering har inte gjorts på senare tid.

Att kontinuerligt revidera och uppdatera samhällsorden i Lexin är nödvändigt för att innehållet ska vara aktuellt och relevant (Se bl.a. Loenheim 2012:302, Loenheim & Hult 2020:50, Höglund & Malmgård 2022:32). Då det finns begränsade medel för Lexin ämnar redaktionen på Språkrådet revidera och uppdatera åtminstone samhällsorden, även om förhoppningen finns att kunna revidera och uppdatera hela ordbasen med jämna intervall.

Med tiden har emellertid den ursprungliga uppmärkning av innehållet i orddatabasen med referens till källtexter som gjordes (Gellerstam 1978:26–27) gått förlorad och det är därför inte helt enkelt att avgöra vilka ord som är samhällsord. Eftersom samhällsorden inte är uppmärkta i Lexins orddatabas eller angivna i de xml-format som används i dag behöver de först och främst identifieras, varefter en revidering och uppdatering med aktuella nya ord kan genomföras. Som ett första steg i arbetet har vi genomfört den här studien.

1.2. Syfte

Syftet med denna artikel är att redogöra för resultaten av vår studie, i vilken vi undersökte möjliga metoder för att identifiera och uppdatera samhällsorden. Vi undersökte även möjligheten att manuellt kategorisera samhällsorden (tematiskt) och märka upp dem för att möjliggöra automatiserad redigering av en delmängd i taget och därmed smidigare och effektivare uppdateringar i framtiden.

2. Samhällsorden

Förutom rent frekvensbaserade skriftspråkliga ord kom Lexins orddatabas att bestå av dels ett ”praktiskt” ordförråd hämtat ur ordindex till ett antal läromedel för invandrarundervisning, diverse vardagliga och talspråkliga ord, ”svåra” ord enligt förståelseundersökningar samt ”samhällsord, dvs. ord som behövs för att man som medborgare skall kunna orientera sig i det svenska samhället (*moms, daghem, kommunalval* etc.)”. (Se Gellerstam 1999:6.)

Samhällsorden i Lexin rör huvudsakligen offentlig sektor (Gellerstam 1999:5) samt kulturspecifika företeelser som *allemansrätt*. Urvalet bygger dels på Ordlista för tolkar (Beijer & Fre Woldu 1997:9), vilken innehåller fackord och samhällstermer inom rättsväsende, allmän försäkring, sjukvård och arbetsmarknad, och dels ingår ”ett tusental ord speciellt utvalda för detta ändamål” (Beijer & Fre Woldu 1997:9). Gellerstam (1978:27) ger exempel på källtexter ur vilka ord inom samhällsinformation till Lexins orddatabas bl.a. har hämtats, titlar som Språkklyftan, Ord vi behöver, Samhälle 1, 2 & 3 samt Socialkunskap (1978:26). I förteckningen över källor till Lexinprojektet finns även ”Tolkordlista” med (1978:26) och bidrog enligt denna med 8 424 uppslagsord, vilket motsvarar 18 % av uppslagsorden i det ursprungliga (obearbetade) källmaterialet.

Det ursprungliga urvalet av samhällsord utarbetades på uppdrag av Invandrarverket och kompletterades till den andra upplagan av Institutionen för svenska språket vid Göteborgs universitet. Samhällsorden försågs med sakupplysningar, förklaringar, och reviderades ”under åren beroende på de fortlöpande ändringar i den verklighet som orden beskriver” (Beijer & Fre Woldu 1997:9). Medvetenheten fanns att man skulle behöva göra systematiska genomgångar för att fånga upp nya samhällsföreteelser, men även om detta sedermera gjordes ”till stor del i samband med översättningsarbetet” (Beijer & Fre Woldu 1997:10) är det oklart när samhällsorden systematiskt reviderats och uppdaterats på senare tid.

Gellerstam (1999:6) beskriver samhällsorden som ”ord för företeelser i det svenska samhället som inte har motsvarigheter i andra kulturer” och tillägger att en exakt sammanvägning av dessa olika typer av ord är svårgörig. Dahl (1999:49) menar att samhällsorden snarast uppfattas som ”ord med utförligare förklaring och sakupplysning”, dvs. deras anknytning till det svenska samhället är inte entydig. Samhällsord återfinns enligt Dahl dels i ord som är ”namnliknande”, dels som en kompletterande upplysning till ord som också ges en vanlig ordförklaring, exempelvis *barnbidrag*. Dahl argumenterar för en snävare definition av vad som räknas som samhällsord i Lexin och menar att typiskt svenska institutioner som *Korpen* hör dit, liksom maträtter som *köttbullar* och *strömming*, men att *addition* inte borde räknas, eftersom samhällsorden bör ha en förankring i den (specifikt) svenska verkligheten. Han påpekar att inte heller kemiska ämnen eller jordarter kan uppfattas som specifikt svenska. Typiska exempel på termer med stark nutidsanknytning till invandrares verklighet är, enligt Dahl, förkortningar från arbetslivet/arbetsförmedlingen.

3. Metod

Vi har identifierat tre olika grundläggande tillvägagångssätt för att ta reda på vilka samhällsorden i Lexin är och vilka ord som inte ingår i Lexin men som kan vara aktuella att klassa som samhällsord och inkludera i Lexin:

- utgå från en lista med kända samhällsord
- utgå från en definition
- identifiera ord som förekommer oftare i viss typ av text än i ”text i allmänhet” (Detta tillvägagångssätt använder vi oss inte av i den här studien.)

Dessa tillvägagångssätt producerar vad vi kan kalla för råmaterial för studier, vilket sedan måste gås igenom manuellt.

Vi har valt att göra tre olika delstudier, där de två första är manuella, medan den tredje använder språkteknologiska metoder. Delstudie ett utgår från en lista med kända samhällsord och använder kategorisering för att identifiera ord som förekommer i samma kontext och påvisa luckor som nya ord kan fylla. Delstudie två undersöker tänkbara samhällsordskandidater utifrån en definition, och den tredje använder en lista med kända samhällsord för att automatiskt ta fram kandidater och med hjälp av definitionen avgöra vilka av dessa kandidater som är rimliga.

3.1. Delstudie 1: Kategorisering

Den här delstudien syftar till att undersöka hur samhällsorden fördelar sig över olika kategorier av ord (ämnesområden) och huruvida en kategorisering kan synliggöra eventuella luckor i samhällsordsbeståndet. Listan som användes var alla ord (lemman) i xml-filen för *Lexins svenska lexikon (Svenska ord)*, version 3 (hädanefter LSL 3) som har försetts med fältet (xml-taggen) "Explanation". Detta gjorde det möjligt att använda en automatiserad sökning och identifiering av samhällsorden. Ord som har tagits bort i LSL 4 togs bort från listan, och alla ord i LSL 4 som har fältet (xml-taggen) "Explanation" lades till.

Eftersom urvalet till samhällsorden i Lexin gjordes till stor del utifrån Ordlista för tolkar använde vi oss av den befintliga indelningen och kategorierna i dessa: *Arbetsmarknad och pension*, *Rättsväsende*, *Socialförsäkring* samt kategorin *Diverse*. Den kliniskt orienterade kategorin *Medicin och psykiatri* i Ordlista för tolkar anpassade vi dock till det mer allmänspråkliga innehållet i Lexin genom att benämna den *Hälsa och sjukvård*.

Vi jämförde ordbasen för Ordlista för tolkar (Tolkordbasen) med vår lista över samhällsord (samhällsordlistan) för att undersöka hur många ord som är direkt hämtade ur Tolkordbasen och i hur stor utsträckning vi kan använda kategorierna ur den. Samtliga ord i samhällsordlistan som även återfanns i Tolkordbasen tilldelades samma kategoriindelning.

3.2. Delstudie 2: Definitionsbaserad undersökning

Denna delstudie syftar till att undersöka hur en definition kan användas för att hitta nya samhällsord. Först enades vi om en definition av samhällsord för denna studie: ”Ord som definierar eller beskriver det svenska samhället, samhällsinstitutioner och företeelser i det, eller som är viktiga i samhällsdebatten, kulturyttringar och traditioner.”

Vi valde ut 200 ord slumpmässigt av de 4 609 ord som tillkommit i LSL 4 i förhållande till LSL 3. Dessa annoterades av två personer var för sig genom att markera om ordet uppfyllde definitionen av samhällsord eller inte. För att mäta samstämmigheten i annoteringen beräknades Cohens kappa (Landis & Koch 1977).

3.3. Delstudie 3: Utökning av lista

Denna delstudie syftar till att med en befintlig lista av samhällsord som grund se om det är möjligt att automatiskt identifiera nya kandidater till samhällsord. Listan som användes var alla ord (lemman) i LSL 3 som har fältet (xml-taggen) ”Explanation”. Ord som har tagits bort i LSL 4 togs bort från listan, och alla ord i LSL 4 som har fältet (xml-taggen) ”Explanation” lades till. Eftersom Lexin använder presens som verbuppslagsform (Malmgren 1999:4), vilket motiveras av att man lättare kan generera övriga verbformer utifrån presens än utifrån infinitiv, anpassades orden så att de överensstämde med den traditionella lemmaformen (infinitiv för

verb). Dessa 465 ord benämndes ”seed-ord” (Riloff & Shepherd 1997, Ahlthorp et al. 2016).

För att tillverka ett ordrum (vektorrum) baserat på distributionell semantik (se till exempel Forsberg & Sköldberg 2022) användes korpussamlingen SVT nyheter (2004–2023), 240 407 472 tokens, och korpusen 8 Sidor, 4 998 634 tokens, totalt ungefär 245 miljoner tokens (hädanefter kallat ”textmaterialet”) (Språkbanken Text 2024a, 2024b). Ordrum baserade på distributionell semantik placerar paradigmatiskt lika ord nära varandra beroende på vilka ord de samförekommer med.

Ordrummet vi tillverkade var av typen Word2Vec (Le & Mikolov 2014) med 100 dimensioner, med en lägsta frekvens på 20 förekomster i textmaterialet och en fönsterstorlek på 5 tokens, det vill säga att algoritmen tar hänsyn till 5 tokens före och 5 tokens efter varje undersökt token i en mening.

Orden samlades in i sin lemmatiserade form, närmare bestämt attributet ”lemma” i Korps XML-filer. Om lemma inte fanns tillgängligt för ordet användes ytformen. Vissa ord har kolon följt av ett tal i lemmaattributet, detta rensades bort innan vidare bearbetning. Från korpusen 8 Sidor rensades meningar som bara bestod av ”8 SIDOR” och ”8 SIDOR / TT” bort.

Ordvektorrummet reducerades från de ursprungliga 100 dimensionerna först till 40 dimensioner med en linjär metod för dimensionalitetsreducering, principalkomponentanalys, och sedan till två dimensioner med t-SNE, en icke-linjär dimensionalitetsreduceringsmetod som förutom att reducera dimensionerna även i viss mån grupperar vektorerna (van der Maaten & Hinton 2008).

För varje seed-ord hittades motsvarande ordvektor, en ”seed-vektor”. Kandidater för nya samhällsord valdes sedan ut genom att för varje vektor bland seed-vektorerna hitta de närmaste 10 vektorerna i det reducerade ordrummet.

Kandidaterna sorterades enligt avståndet till den närmaste seed-vektorn, och de 100 kandidaterna med kortast avstånd val-

des ut för utvärdering. Utvärderingen skedde genom att välja ut 100 slumpmässigt valda ord från de ord i textmaterialet som hade minst 20 förekomster och blanda dessa med de 100 kandidaterna. Inga av de slumpmässigt valda orden råkade bli någon av kandidaterna. De totalt 200 orden presenterades i en slumpmässig ordning för två annoterare (samma ordning för båda annoterarna), och annoterarna fick samma uppgift som ovan, att markera om ordet uppfyllde definitionen av samhällsord eller inte. För att mäta samstämmigheten i annoteringen beräknades Cohens kapp (Landis & Koch 1977).

Listan med de 200 orden gicks igenom av båda annoterarna och varje ord som någon hade markerat diskuterades och en ensad uppfattning antecknades utan att annoterarna visste vilka ord som var kandidater och vilka som var slumpmässigt utvalda. Efter ensningen delades orden upp i kandidater och slumpmässigt utvalda. Ett χ^2 -test (Campbell, David & Walters 2007:132) genomfördes för att se om kandidaterna hade signifikant fler samhällsord än de slumpmässigt utvalda.

4. Resultat

4.1. Delstudie 1: Kategorisering

Av totalt 465 ord i samhällsordlistan fanns 234 ord (50 %) även i Tolkordbasen och indelades därmed i samma kategorier. Vid en sortering av listan fick vi en överblick över mängden ord per kategori för de ord som även ingick i Tolkordbasen: Arbetsmarknad och pension 68 ord, Diverse 14 ord, Hälsa och sjukvård 28 ord, Rättsväsende 70 ord, Socialförsäkring 54 ord.

231 ord fanns inte med i Tolkordbasen. Dessa indelades i kategorier utifrån semantiskt innehåll. Vi tillämpade så långt det

var möjligt samma kategorisering som i Tolkordbasen, med viss försiktighet och i förekommande fall semantisk granskning. Vid denna indelning utökades kategorin Diverse i Lexin från 14 ord till 91, varför en indelning av givna ord även i alternativa (under-) kategorier, exempelvis Sveriges landskap, kan vara till nytta.

Delstudie 1 visade att kategorier kan vara användbara för att synliggöra luckor i en samling av utvalda ord, vilka kan utgöras av såväl luckor (saknade ämnesområden) bland de kategorier som ingår i listan och luckor (saknade ord) inom en viss kategori av ord. Exempelvis återfanns såväl *valborgsmässoafton* (se figur 2) som *fruntimmersveckan* i listan över samhällsord, men inte *mid-sommar*.

The screenshot shows the Lexin website interface. At the top, there is a dark teal header with the 'LEXIN' logo on the left, a language dropdown menu set to 'svenska' in the center, and a hamburger menu icon on the right. Below the header is a search bar containing the text 'valborgsmässoafton' with a magnifying glass icon to its right. The search results are displayed in a light blue box. The main entry is 'valborgsmässo | afton' with its phonetic transcription '[²və:lbårjsmes:oaf:tån]'. Below this, it is identified as a 'substantiv' and lists its plural forms: '(valborgsmässo | afton · valborgsmässoaftonen · valborgsmässoaftnar · valborgsmässoaftnarna)'. To the right of the plural forms is a speaker icon. Below the light blue box, the text 'den 30 april (som anses symbolisera vårens ankomst)' is shown. Further down, under the heading 'Exempel:', there is a bulleted list: 'de samlades runt bålet på valborgsmässoafton och sjöng vårsånger'.

Figur 2: Artikeln för ordet valborgsmässoafton på lexin.se, vilken har försetts med extra information inom parentes, men ingen specifik förklaring.

Kategorisering visade sig även användbar för att synliggöra regelbundenheter och inkonsekvenser vad gäller redaktionell hantering. I användargränssnittet har vissa av orden getts såväl definition som förklaring och andra har försetts med enbart en förklaring.

Lemmat *Högsta domstolen* har exempelvis såväl en definition som förklaring (där även ordet ”Förklaring:” har angetts), vilket visas i figur 3. Jämför vidare orden *inkomstpension*, som har försetts med förklaring, och *ålderspension*, där ingen förklaring har angetts.



Figur 3: Ordet *Högsta domstolen* på lexin.se har försetts med både definition och förklaring.

4.2. Delstudie 2: Annotering av 200 nya ord ur LSL 4

Av de 200 undersökta orden i delstudie 2 var de två annoterarna överens om att 9 av orden uppfyllde den överenskomna definitionen av samhällsord. För ytterligare 17 ord var annoterarna oense om ordet uppfyllde definitionen. Samstämmigheten enligt Cohens kapp beräknades till 0,47 (måttlig enligt Landis & Koch 1977), se tabell 1.

	Ja	Nej
Ja	9	5
Nej	12	173

Tabell 1: Annoteringen av de 200 nya orden ur LSL 4 gav en samstämmighet på 0,47 (Cohens kapp, måttlig enligt Landis & Koch 1977).

4.3. Delstudie 3: Utökning av lista

I delstudie 3 förekom 129 438 ord minst 20 gånger i textmaterialet (SVT nyheter + 8 Sidor) och utgjorde därmed underlaget för ordrummet. Bland dessa (hög)frekventa ord fanns 392 av seed-orden. Av de totalt 465 seed-ord som identifierades i befintliga Lexin kunde 73 seed-ord alltså inte användas i framtagandet av kandidater, eftersom de inte fanns med i textmaterialet minst 20 gånger (och därmed beräknades inga vektorer för dem). Resultatet av reduceringen till två dimensioner kan ses i figur 4, där varje ordvektor representeras av en punkt i figuren. Seed-vektorerna är märkta med mörkgul färg och kandidaterna med turkos färg i figur 4. En detaljbild med ord relaterade till fackföreningar finns i figur 5.

Av de 200 orden som valts ut för annotering var de två annoterarna överens om att 35 av orden uppfyllde definitionen av samhällsord. För ytterligare 38 ord var annoterarna oense om ordet uppfyllde definitionen. Samstämmigheten enligt Cohens kapp beräknades till 0,53 (måttlig enligt Landis & Koch 1977), se tabell 2.

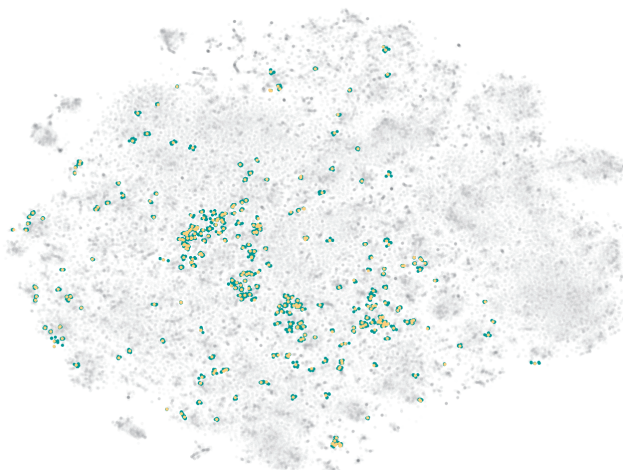
	Ja	Nej
Ja	35	8
Nej	30	127

Tabell 2: Annoteringen av 100 kandidater och 100 slumpmässiga ord (totalt 200 ord) gav en samstämmighet på 0,53 (Cohens kapp, måttlig enligt Landis & Koch 1977).

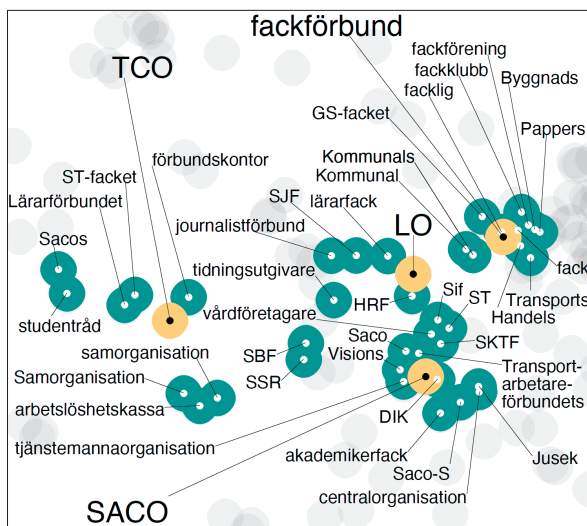
Ensningdiskussionen gav även upphov till att några ord som ingen hade markerat blev markerade vid ensningen.

	Kandidater	Slumpmässiga ord
Samhällsord	57	14
Inte samhällsord	43	86

Tabell 3: Resultat efter ensning.



Figur 4: Alla ord som förekommer minst 20 gånger i textmaterialet, med seed-vektorer märkta med mörkgul färg och kandidaterna märkta med turkos färg.



Figur 5: Detalj som visar ord som relaterar till fackföreningar, med seed-vektorer märkta med mörkgul färg (centrum visas med en svart prick) och kandidaterna märkta med turkos färg (centrum visas med en vit prick). Ord som inte är kandidater visas med en något mörkare grå i förhållande till figur 4.

Resultatet av den ensade annoteringen visas i tabell 3, där 57 kandidater av 100 ansågs vara samhällsord. En fullständig lista på dessa 57 kandidater finns i tabell 4.

Chi2-testet gav resultatet 38,52, vilket gör att vi kan förkasta nollhypotesen att skillnaden mellan kandidaternas 57 % bedömda samhällsord och de slumpmässiga ordens 14 % bedömda samhällsord bara beror på slump med en sannolikhet $p < 0.001$. Skillnaden mellan listan av kandidater och listan av slumpmässiga ord är alltså signifikant.

socialbidrag	delbetänkande
Bris	april-firande
vårdboende	länstyrelse
Region	Östersjölandskapen
hockeyallsvenska	elevkår
kallelse	miljöparti
Elitserien	V
prövningstillstånd	regionval
Sjuhäradsbygden	förvaltningsrätt
f-kassa	steroid
polisstyrelse	pastorat
Gävleborg	påskdag
nyårsdag	riksbank
a-kassa	hetvägg
arbetsmiljöverk	juldag
midsommardag	allsvenska
vänsterparti	anhöriginvandring
byggrätt	kammarrätt
universitetsadjunkt	Sörmland
Västernorrland	ob-tillägg
betänkande	Nordsjön

registerutdrag	Naturvårdsverk
brottmål	brottmål
skatteväxling	arbetsbrist
europaparlamentsval	advokatfirma
allhelgonahelg	familjeåterförening
påskafton	studiestöd
valkuvert	sörmland
sjuklön	

Tabell 4: De 57 automatiskt framtagna kandidater till samhällsord där annoterarnas konsensus var att ordet skulle räknas som samhällsord.

5. Diskussion

Delstudie 1 visade att en indelning i kategorier kan vara användbar för mer frekventa och tidsmässigt hanterbara framtida deluppdateringar genom att man kan revidera och uppdatera en eller några kategorier i taget. Vi fann att det vore lämpligt att märka upp på det här sättet och när det är möjligt använda samma kategorier som i Ordlista för tolkar.

För att möjliggöra detta bör man vara konsekvent med att lägga till kategorier i xml vid revidering och uppdatering, samt därvid tillämpa ett konstant regelverk. Möjlig märkning kan vara att lägga till ett attribut på taggen "Explanation" alternativt på taggen för lemmat samt döpa attributet (exempelvis "Category"). Vissa ord kan tillhöra flera olika kategorier. Det är därför en god idé att redan från början bestämma huruvida man vill tillåta indelning i flera kategorier, åtskilda med kommatecken exempelvis. Man bör givetvis vara konsekvent och tillämpa samma principer vid kategorisering och märkning även över tid.

Både i delstudie 2 och 3 var samstämmigheten mellan annoterarna måttlig. Samstämmigheten kan antagligen förbättras genom att förfina definitionen av vad som utgör samhällsord (se avsnitt

3.2) och genom löpande diskussioner mellan annoterarna om gränsdragningar. Var gränsen går mellan vad som är samhällsord och vad som är fackord eller termer inom något av områdena politik, ekonomi eller juridik är inte alltid helt lätt att avgöra.

Delstudie 3 visade att det var signifikant mer effektivt att leta samhällsord i en automatiskt extraherad lista baserad på närhet i ett ordrum än att gå igenom slumpmässigt valda ord. Det är därför ett sätt att få tag i så många samhällsord som möjligt, men däremot inget säkert sätt att få en uttömmande lista. Där kan en lista indelad i kategorier utgöra ett bra komplement genom att visa på luckor inom en viss ordkategori eller domän och därmed möjliggöra analys och fokuserad (manuell eller automatiserad) uppdatering av orden i en eller några kategorier åt gången. Begränsad arbetsinsats krävs för att uppdatera en avgränsad del av listan och gör det därmed möjligt att mer regelbundet och effektivt uppdatera samhällsorden i Lexin i framtiden.

Litteratur

Ordböcker och korpusar

- LSL 3 = *Lexins svenska lexikon*, version 3 [Nätversionen av] *Svenska ord: med uttal och förklaringar* (1995). 3 uppl. Stockholm: Myndigheten för skolutveckling/Liber.
- LSL 4 = *Lexins svenska lexikon*, version 4 [Nätversionen av] *Svenska ord* (2011). 4 uppl. Språkrådet, Institutet för språk och folkminnen.
- Lexin svensk-engelskt lexikon* (2004). Stockholm: Myndigheten för skolutveckling.
- Språkbanken Text (2024a). SVT nyheter [Data set]. Språkbanken Text. <<https://doi.org/10.23695/vvd5-j8482>>.

Språkbanken Text (2024b). 8 Sidor (Uppdaterad: 2024-03-07) [Data set]. Språkbanken Text. <<https://doi.org/10.23695/f6ds-fo45>>.

Annan litteratur

- Ahltorp, Magnus, Maria Skeppstedt, Shiho Kitajima, Aron Henriksson, Rafal Rzepka & Kenji Araki (2016): Expansion of medical vocabularies using distributional semantics on Japanese patient blogs. I: *Journal of Biomedical Semantics* 7 (58), 1–18. <<https://doi.org/10.1186/s13326-016-0093-x>>.
- Beijer, Mai & Kiros Fre Woldu (1997): *Detta är LEXIN. Lexikon för invandrare. Lägesbeskrivning våren 1997*. Stockholm: Skolverket.
- Campbell, Michael J., David Machin & Stephen J. Walters (2007): *Medical Statistics: a Textbook for the Health Sciences*. 4. ed. Chichester: Wiley.
- Dahl, Adolf (1999): Lexinlexikonen och översättaren. I: *LexicoNordica* 6, 47–53.
- Forsberg, Markus & Emma Sköldberg (2022): Ordvektorer i lexicografiskt arbete. I: Elena Volodina, Dana Dannélls, Aleksandrs Berdicevskis, Markus Forsberg & Shafqat Virk (eds.): *LIVE and LEARN – Festschrift in honor of Lars Borin*. Gothenburg: Department of Swedish, Multilingualism & Language Technology. 46–50.
- Gellerstam, Martin (1999): LEXIN – Lexikon för invandrare. I: *LexicoNordica* 6, 3–18.
- Gellerstam, Martin (1978): Att välja sina ord: om ordurval till enspråkiga ordböcker. *Rapporter från Språkdata* 9. Göteborgs universitet, Institutionen för språklig databehandling.
- Höglund, Laila & Sofia Malmgård (2022): *Framtidens Lexin – forskningsöversikt*. Rapporter från Språkrådet 13. Institutet för språk och folkminnen.

- Landis, J. Richard & Gary G. Koch (1977): The Measurement of Observer Agreement for Categorical Data. I: *Biometrics* 33(1), 159–174. <<https://doi.org/10.2307/2529310>>.
- Loenheim, Lisa & Ann-Kristin Hult (2020): *Framtidens Lexin – forskningsöversikt*. Rapporter från Språkrådet 12. Institutet för språk och folkminnen.
- Malmgren, Sven-Göran (2012): Från Svenska ord 3 (Lexin) till Svenska ord 4. I: Birgit Eaker, Lennart Larsson & Anki Mattisson (red.): *Nordiska studier i lexikografi* 11. *Rapport från Konferensen om lexikografi i Norden, Lund 24–27 maj 2011*. 454–465.
- Malmgren, Sven-Göran (1999): De svenska LEXIN-ordböckerna ur ett (ställföreträdande) användarperspektiv. I: *LexicoNordica* 6, 79–90.
- van der Maaten, Laurens & Geoffrey Hinton (2008): Visualizing data using t-SNE. I: *Journal of Machine Learning Research* 9(86), 2579–2605.
- Le, Quoc & Tomas Mikolov (2014): Distributed representations of sentences and documents. I: *Proceedings of the 31st International Conference on Machine Learning (ICML ‘14)* 32. 1188–1196.
- Riloff, Ellen & Jessica Shepherd (1997): A Corpus-Based Approach for Building Semantic Lexicons. I: *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing (EMNLP-2)*. 117–124.
- Svensén, Bo (2004): *Handbok i lexikografi. Ordböcker och ordboksarbete i teori och praktik*. 2 uppl. Stockholm: Norstedts Akademiska Förlag.
- Tarp, Sven (2008): *Lexicography in the Borderland between Knowledge and Non-Knowledge. General Lexicographical Theory with Particular Focus on Learner’s Lexicography*. Lexicographica Series Maior 34. Tübingen: Max Niemeyer Verlag.

Tarp, Sven (1999): Lørnerordbøger for indvandrere og andet godtfolk. I: *LexicoNordica* 6, 107–132.

Magnus Ahltop
språkteknolog
Nakajima Koen Research Institute
Hildebergsvägen 26
SE-117 62 Stockholm
magnus@nakajimakoen.org

Laila Höglund
projektledare för Lexin
Språkrådet, Institutet för språk och
folkminnen
Alsnögatan 11
SE-116 41 Stockholm
laila.hoglund@isof.se