Nordiska studier i lexikografi

16

Rapport från
16:e konferensen om lexikografi i Norden
Lund 27–29 april 2022

REDIGERAD AV
Louise Holmer
Greta Horn
Hans Landqvist
Pär Nilsson
Eva Nordgren
Emma Sköldberg

Skrifter utgivna av Nordiska föreningen för lexikografi
Skrift nr 17
i samarbete med Meijerbergs institut för svensk etymologisk forskning

Nordiska studier i lexikografi 16 rapporterar från den 16:e konferensen i lexikografi, som genomfördes i Lund 27–29 april 2022.

Volymen innehåller 30 bidrag som bygger på inlägg från konferensen i form av plenarföreläsningar, sektionsföredrag och posterpresentationer. Artiklarna spänner innehållsligt över ett brett fält, men samtliga anlägger någon form av lexikografiskt perspektiv. Flera av dem anknyter till konferensens tema *Lexikografiska utmaningar.*

Merparten av bidragen är författade på danska, norska eller svenska, men ett mindre antal är skrivna på engelska.

NFL NORDISKA FÖRENINGEN FÖR LEXIKOGRAFI

NORDISK FORENING FOR LEKSIKOGRAFI • LEKSIKOGRAFIAN POHJOISMAINEN YHDISTYS
NORÐURLENDSKT FELAG FYRI ORDABÓKAGERD • NUNANI AVANNARLERNI
OQAASERSIORFILIORTUT PEQATIGIIFFIAT • NORRÆNA ORÐABÓKARFRÆÐIFÉLAGIÐ
NORDISK FORENING FOR LEKSIKOGRAFI • DAVVIRIIKALAŠ LEKSIKOGRÁFALAŠ SEARVI
NORDISKA FÖRENINGEN FÖR LEXIKOGRAFI

Nordiska studier i lexikografi
16

Nordiska studier i lexikografi 16

Rapport från 16:e konferensen
om lexikografi i Norden

Lund 27–29 april 2022

Redigerad av
Louise Holmer
Greta Horn
Hans Landqvist
Pär Nilsson
Eva Nordgren
Emma Sköldberg

Skrifter utgivna av Nordiska föreningen för lexikografi
Skrift nr 17

Meijerbergs arkiv för svensk ordforskning
48

Lund och Göteborg 2023

Nordiska studier i lexikografi 16

Ingår som del 48 i Meijerbergs arkiv för svensk ordforskning

Redaktion:
Louise Holmer
Greta Horn
Hans Landqvist
Pär Nilsson
Eva Nordgren
Emma Sköldberg

Tryck: By Wind
Sättning samt formgivning av bokens omslag: Jocke Wester

ISBN: 978-91-986791-5-1 (tryck)
ISBN: 978-91-986791-6-8 (digital)
ISSN: 1894-4663 (Nordiske studier i leksikografi, tryck)
ISSN: 2246-7823 (Nordiske studier i leksikografi, online)
ISSN: 0803-9313 (Skrifter udgivet af Nordisk Forening for Leksikografi)
ISSN: 0348-7741 (Meijerbergs arkiv för svensk ordforskning)

Utgiven med stöd från:

Nordplus

GÖTEBORGS UNIVERSITET

MEIJERBERGS INSTITUT

Svenska Akademien

NFL Nordiska föreningen för lexikografi

Innehåll

Förord

Den 16:e konferensen om lexikografi i Norden hölls onsdag till fredag 27–29 april 2022 i Lund i universitetets lokaler. Konferensen var ett samarrangemang mellan Nordiska föreningen för lexikografi (NFL), Svenska Akademiens ordboksredaktion och Institutionen för svenska, flerspråkighet och språkteknologi vid Göteborgs universitet.[1] Partnerinstitutioner var dessutom Universitetet i Bergen och Árni Magnússon-institutet för isländska studier i Reykjavík.

De återkommande konferenserna i NFL:s regi är de största och viktigaste mötestillfällena för aktiva lexikografer och övriga lexikografiintresserade i de nordiska länderna. Med tanke på att ordboksmiljöerna är så små i de enskilda länderna är dessa nordiska möten ytterst värdefulla.

Temat för konferensen var *Lexikografiska utmaningar*. Med temat ville arrangörerna uppmuntra till presentationer med stor ämnesmässig spridning och bredd. Målet var att forskare, aktiva lexikografer, översättare, pedagoger m.fl. skulle mötas och diskutera den lexikografiska situationen i Norden och att genom konferensinläggen få nya perspektiv på de utmaningar som är knutna till samtida ordboksframställning.

Förutom det vetenskapliga programmet med plenarföredrag, föredrag i parallella sektioner och en särskild postersektion, innehöll konferensen sociala programpunkter. En välkomstmottagning anordnades på tisdagen, kvällen före konferensens första dag, i Svenska Akademiens ordboksredaktions lokaler. Onsdagseftermiddagen avslutades med en stadsrundvandring och en guidad visning av Skissernas museum i Lund. Konferensmiddagen hölls på torsdagskvällen i Akademiska Föreningens borg i Lundagård till tonerna av storbandsjazz.

Konferensen lockade sammanlagt drygt 100 deltagare från Belgien, Danmark, Finland, Färöarna, Island, Norge, Serbien, Sverige och Tyskland. Ett antal av bidragen anslöt sig till temat. I dessa behandlades t.ex.

1 NFL:s konferenser har arrangerats vartannat år sedan starten 1991. Med anledning av covidpandemin och myndigheternas då gällande restriktioner kunde konferensen dock inte hållas som planerat år 2021. I stället genomfördes den med samtliga deltagare på plats i Lund våren 2022.

utmaningar med teckenspråkslexikografi, automatisk termextraktion och ordböcker för skriftlösa språk.

Tre inbjudna plenarföreläsare inledde konferensdagarna med var sitt föredrag:

- Professor Dirk Geeraerts, KU Leuven, Belgien: *Does the data deluge the dictionary?*
- Professor Lisa Holm, Lunds universitet: *Lexikal verbsemantik i tre dimensioner – en möjlig ingång till lexikografisk beskrivning?*
- Professor Dr. Carolin Müller-Spitzer, Universität Mannheim, Tyskland: *The user in focus – what does that really mean?*

Totalt omfattade konferensen drygt 50 bidrag i form av föredrag och posterpresentationer. Ett flertal av dessa har därefter arbetats om till skriftliga inlägg. Konferensvolymen innehåller sammanlagt 30 bidrag. Huvuddelen av dem är skrivna på ett av de nordiska språken danska, norska eller svenska, men ett mindre antal är skrivna på engelska.

Varje bidrag har granskats av två anonyma granskare och dessutom av en redaktör från redaktionskommittén.

De skriftliga bidragen förväntas resultera i nya kunskaper som kan fördjupa och förnya det lexikografiska arbete som bedrivs vid forskningsinstitutioner, förlag och företag samt genom crowdsourcing i hela Norden.

Konferensen har fått bidrag från Nordplus Nordens språk, från Nordiska föreningen för lexikografi och från Svenska Akademien. Dessutom har stöd mottagits från Göteborgs universitet samt från Meijerbergs institut för svensk etymologisk forskning, detta för färdigställandet av konferensvolymen. Arrangörerna är ytterst tacksamma för detta stöd.

Redaktionskommittén har bestått av Louise Holmer, Greta Horn, Hans Landqvist och Emma Sköldberg från Göteborgs universitet samt Pär Nilsson och Eva Nordgren från SAOB-redaktionen.

Redaktörerna tackar alla som bidragit till att konferensen kunde arrangeras och att volymen har kunnat färdigställas och publiceras.

Lund och Göteborg i december 2023

Louise Holmer, Greta Horn, Hans Landqvist,
Pär Nilsson, Eva Nordgren och Emma Sköldberg

Plenarföreläsning

Lexikal verbsemantik i tre dimensioner

Lisa Holm

This article suggests a model for lexical verb semantics that unifies Halliday's model for transitivity with Langacker's concept of base and profile and also with a model for situation type (cf. Smith 1997). It is claimed that verbs adhere to one of four macro domains: the material, the social, the mental or the relational domain. Within these domains there are numerous types of verbs, such as motion verbs, production verbs and speech act verbs. Verbs with the same base or micro domain differ in lexical profile. The Swedish counterparts to *stab* and *kill* belong to a domain of "violence afflicted to human beings"; here *kill* is a transitional verb while *stab* is instrumental. It is argued that among material verbs there are both verbs with material profiles, such as manner verbs, and verbs with other types of profiles. Verbs like *run*, *laugh* and *shiver* describe how a referent is involved in some specific physical constellation. Activities like *work*, *play* and *cook* are also performed by human bodies; still these verbs don't indicate exactly how the body is involved. You can work, play or cook in many different ways. When it comes to situation type it is suggested that there are three main lexical types (state verbs, process verbs and transitional verbs) in Swedish, which in different constructions of the verb phrase can express the major situation types: state, activity, accomplishment, achievement and semelfactive. Finally some aspects of role semantics are considered.

KEYWORDS: verb semantics, lexical profile, situation type, role semantics

1. Inledning

Att beskriva verb i en ordbok är en erkänt svår uppgift, särskilt om det rör sig om något av de polysema verb som språket är så rikt på. För lexikografen är det omöjligt att ha en strikt teoretisk infallsvinkel på en sådan uppgift. Det handlar istället om att urskilja huvudbetydelser och underbetydelser, som betingas av verbets konstruktionssätt och kontext, och presentera dem så stringent och brukarvänligt som möjligt. Det räcker att titta i *Svensk ordbok* (SO) eller *Svenska Akademiens ordbok* (SAOB) på verb som *driva*, *luta* och *riva* – inget av dem svenska språkets mest polysema verb – för att inse att verbsemantik är ett mylle. Min avsikt

med denna artikel är därför inte att erbjuda en modell för lexikografisk beskrivning av verb. Syftet är istället att sätta in verbens lexikala semantik i ett teoretiskt sammanhang. Förhoppningen är att beskrivningen ska vara användbar som ett sätt att tänka kring verb också för en lexikograf. Perspektivet här är generellt och utgår från basala verb som inte analyseras på djupet. Det jag försöker göra är att förena tre teoretiska ingångar till verbbetydelse: lexikal typ, aktionsart och semantiska roller. Mest fokus läggs på lexikal typ och, i förlängningen av denna, lexikal profil.

2. Några utgångspunkter

Verbens uppgift i språket är att ange konstellationer och deras olika konstruktionssätt hänger ihop med det. Alltså är syntax och semantik respektive verbsemantik och satssemantik oskiljaktiga. Ändå talar jag här så renodlat som möjligt om semantik och fokuserar på de delar av betydelsen som ligger i eller närmast själva verbet. Jag föreslår följande schematisering av den samlade semantiken kring verb:

Den lexikala semantiken (verb och verbfras)
lexikal typ
valens och semantiska roller
aktionsart
Den finita semantiken (satsen)
aspekt
modus
tempus

FIGUR 1. Verbsemantikens olika nivåer.

Lexikal typ handlar om vilken sorts erfarenhetsmässigt innehåll ett enskilt verb bidrar med. Valens och roller handlar om vilket eller vilka konstruktionsscheman ett verb har, syntaktiskt och semantiskt. Aktionsart handlar om vilka temporala egenskaper verb och verbfraser uppvisar. Relationen mellan aktionsart och aspekt är i ett språk som svenska, som saknar morfologisk aspekt, ytterst svårfångad, men eftersom jag här enbart talar om lexikal typ, aktionsart och semantiska roller utgör det inget problem.

Verb kan ha mycket generell betydelse (som *ha* och *bli*) eller mycket specifik (som *fritera* och *sautera*) och allt däremellan. Olika verb zoomar in på verkligheten i olika grad. Verbens rikliga polysemi hänger sannolikt ihop med både deras grundläggande relationella betydelse och hur pass generella eller specifika de är. Riktigt specifika verb är normalt inte så polysema.

Fyra termer som behövs för att tala om verbsemantik hämtar vi från *Svenska Akademiens grammatik* (SAG 1999): aktion, aktant, animat och inanimat. En aktion är det som ett verb utpekar, t.ex. en handling, en händelse, en process, ett tillstånd eller en relation, alltså något som finns i den verklighet talaren refererar till. En aktant är en referent som har en semantisk roll i förhållande till aktionen. Referenter kan vara animata (levande varelser med förnuft, vilja, känsla) eller inanimata (referenter som inte förutsätts utrustade med förnuft, vilja, känsla). I exemplet *Anna hämtade boken* anger verbet själva kärnan i aktionen 'att hämta något', och nominalfraserna utpekar aktanter som har semantiska roller i förhållande till aktionen, i det här fallet en animat aktant, 'Anna', som utsätter en inanimat aktant, 'boken', för en aktion.

3. Lexikal typ på mellan- och makronivå

I olika sammanhang används etiketter för semantiska grupper av verb, t.ex. rörelseverb (*cykla*, *springa*, *åka*); placeringsverb (*lägga*, *ställa*, *sätta*); positionsverb (*ligga*, *sitta*, *stå*); instrumentverb (*dammsuga*, *kamma*, *skära*); väderverb (*hagla*, *snöa*, *åska*); produktionsverb (*baka*, *bygga*, *skriva*); sägeverb (*berätta*, *fråga*, *säga*); perceptionsverb (*höra*, *märka*, *se*) etc. Sådana etiketter visar att grupper av verb hänger ihop semantiskt men påvisar också en beskrivningsmässig mellannivå av lexikala fält med besläktade verb som vi behöver kunna urskilja. Ändå finns det, vad jag vet, inte någon uttömmande taxonomi över verbtyper på denna nivå för svenska verb, och SAG fastslår inga sådana termer. Likväl finns det värdefulla studier kring några av dem, t.ex. Jakobsson (1996), Viberg (2008) och Kinn et al. (2018).

För att hantera verblexemen mera generellt behöver man också en översiktskarta, dvs. lexikala typer på makronivå. En sådan översikt finns i SAG (kap. 7 (Verb), §11). En liknande modell finns inom den systemisk-funktionella grammatiken (SFG), för engelskans del i Halliday & Matthiessen

(2014) *Introduction to Functional grammar* (kap. 5) och för svenskans del i Holmberg & Karlsson (2006) *Grammatik med betydelse* (kap. 4). Själv har jag kommit fram till en likartad grovskiss genom att göra aktionsartsanalys på drygt 200 centrala svenska verb (Holm u.å.). Beroende på vilka likheter och skillnader mellan olika verbtyper man vill betona eller bortse ifrån kan man teckna kartan på olika sätt. Min version ges i figur 2. (Termmässigt ansluter jag här främst till SFG.)

I. Det **materiella** fältet	*gräva, rusa, snöa*
II. Det **sociala** fältet	*berätta, uppfostra, överlämna*
III. Det **mentala** fältet	*förakta, förvänta, inbilla sig*
IV. Det **relationella** fältet	*vara, utgöra, äga*

FIGUR 2. Verbala betydelsefält på makronivå.

Verben i det materiella fältet används för att beskriva vad som sker i den fysiska världen, både av sig själv och för att människor agerar i den: *fladdra*, *glufsa*, *guppa*, *hosta*, *hämta*, *krokna*, *regna*, *skriva*, *springa*, *städa*. En stor andel av verben i det materiella fältet är typiska människoverb, men här finns också djur-, natur- och fysikverb m.fl. Barn lär sig materiella verb först och bygger upp sin grammatik kring dem (jfr Christensen 2010). Man kan våga anta att det prototypiska verbet är ett materiellt verb. Verben i det sociala fältet handlar om vårt sociala agerande och kommunicerande: *ge*, *hjälpa*, *leka*, *samarbeta*, *trösta*; *avslöja*, *nominera*, *ropa*, *säga*, *viga*. Verben i det mentala fältet beskriver vad som äger rum inom oss, tänkande och kännande: *gilla*, *hata*, *mena*, *tro*, *tycka*, *tänka*, *veta*, *älska*. Verben i de sociala och mentala fälten är förstås utpräglade människoverb. Verben i det relationella fältet används för att beskriva världen mera objektivt, utan människokroppen som utgångspunkt, med olika egenskaper, etiketter och relationer: *bestå av*, *beteckna*, *bli*, *dominera*, *ha*, *heta*, *likna*, *tåla*.

Inom alla makrofält finns undergrupper av olika typer (jfr etiketterna ovan), men bara det sociala fältet sönderfaller tydligt i två delar: 1) socialt handlande och 2) språklig kommunikation. Verbet *ge ngn ngt* är centralt i den första gruppen och *säga ngn ngt* i den andra. Verb för socialt handlande har annars många likheter med materiella verb, men de bör urskiljas som egen grupp för att de utgör en stor och kommunikativt viktig grupp.

Verb för språklig kommunikation har många likheter med mentala verb, bland annat för att både sägeverb och mentala verb kan ta *att*-sats: *Hon {påstod/förmodade} att han ljög*.

Grovindelningen tar inte hänsyn till metaforik, men många verb från makrofält I och II fungerar relationellt i metaforisk användning, t.ex. *Vägen går till Malmö*. Den tar heller inte hänsyn till annan polysemi utan försöker utgå ifrån den mest prototypiska och basala användningen av varje verb. Indelningen låtsas inte heller om några gränsfall, även om sådana finns mellan alla fält, och inom varje fält finns det mer eller mindre prototypiska medlemmar.

Även om det inte går att dra några skarpa gränser är indelningen meningsfull. Den är ett kraftfullt verktyg vid textanalys, och man behöver sällan göra fullständig processanalys i SFG:s bemärkelse för att få ett bra grepp om en text. Det räcker ofta att konstatera vilka verbtyper som finns i texten och vilka som dominerar. Texter som har ett stort inslag av materiella verb är vanligen berättande. Texter som domineras av relationella verb är beskrivande och utredande. Det är stor skillnad på texter med och utan sägeverb, eftersom sådana öppnar för fler röster än avsändarens. Mentala verb låter läsaren överskrida gränsen mellan yttre och inre och är därför så gott som otänkbara i texter som eftersträvar objektivitet.

4. Aktionsart

En dimension av verbens semantik som bara i viss mån följer de lexikala makrofälten är aktionsart. Aktionsart är en uppsättning temporala grundmönster som verb och verbfraser kan uttrycka. Exakt vilken status dessa mönster har är oklart, men de påverkar vilka hjälpverb och adverbial olika verbfraser kan kombineras med och avgör hur vi tolkar satser temporalt. I litteraturen finns en rik uppsättning aktionsartsmodeller, som betonar olika saker, ofta beroende på vad de ska användas till. Gemensamt för de allra flesta är att de förhåller sig till Vendlers klassiska artikel från 1957 (”Verbs and times”). Något av en standardmodell finns hos Smith (1997) *The Parameter of Aspect*, och det är den jag utgår ifrån här, se figur 3. En tidig svensk version av modellen finns i Christensen (1995); i stort sett samma modell återfinns i SAG kap. 33 (Aktionsarter), även

om de termer jag använder här avviker något från SAG:s.[1] I Croft (2012) sätts modellen in i ett konstruktionsgrammatiskt och kognitivt semantiskt ramverk. Sasse (2002:263) efterlyser i sin ”state-of-the-art”-artikel en modell som tar itu med verbens lexikala roll: ”An urgent desideratum is the investigation of the role of the lexicon, in particular the subcategorization of situation types. It has become clear that Vendler classes do not suffice.” I denna artikel försöker jag ta itu med just detta på en principiell nivå.

Smith (1997)	**Svenska termer**
State	Tillstånd
Activity	Process
Accomplishment	Gränsrelaterad process
Achievement	Tillståndsövergång
Semelfactive	Semelfaktiv

FIGUR 3. En aktionsartsmodell.

Tillstånd är statiska aktioner som inte händer utan bara är eller föreligger: ’ha långt hår’, ’heta Pia’, ’vara arg’, ’äga en cykel’. Processer är dynamiska aktioner; något händer och aktionen är ofta tydligt uppbyggd av olika moment eller steg, mest typiskt sådant som ’andas’, ’plaska’, ’prata’, ’ringla’, ’stappla’, ’studsa runt’. Processer kan bestämmas av tidsadverbial för varaktighet, som ***i*** *tio minuter* (SAG kap. 20 (Verbfraser: adverbial), §104). Processer kan vara gränsrelaterade, t.ex. ’simma till stranden’, ’springa hem’, då finns ett mål för processen, eller ’baka en tårta’, ’regna bort’, då finns någon sorts resultat, och de kan bestämmas av tidsadverbial för tidsåtgång som ***på*** *tio minuter* (SAG kap. 20 (Verbfraser: adverbial), §105). En tillståndsövergång beskriver en övergång från ett tillstånd till ett annat, t.ex. ’stänga’, från öppen till stängd; ’dö’, från levande till icke-levande, och vi brukar uppfatta sådana aktioner som momentana. Semelfaktiver beskriver också något som sker väldigt kortvarigt men utan att förändra utgångsläget, t.ex. ’hicka’, ’nysa’ och ’hoppa till’.

Ett och samma verb kan vanligen uttrycka mer än en aktionsart, beroende på konstruktion och kontext. Det är ett mycket rörligt system, och

1 För termen *gränsrelaterad process*, se Ekberg (1989:101, not 2).

att studera aktionsart är i hög grad en fråga om att studera aktionsartsväxlingar och hitta mönster bland dem. Vissa verb kan genomgå vissa växlingar, men inte alla. Ibland gäller växlingarna stora grupper av verb; ibland är de nästintill individuella. Två generella aktionsartsväxlingar illustreras i exempel (1) och (2).

1 Bo dricker kaffe
a) 'sitter och dricker kaffe' process
b) 'är en kaffedrickande person' tillstånd

Så gott som alla verb som inte är tillståndsverb kan tolkas som något som faktiskt äger rum (1a) eller generiskt/habituellt, dvs. som ett tillstånd som innebär att en aktion brukar eller kan inträffa (1b). Detta är den mest generella aktionsartsväxlingen av alla.

2 My hickade
a) 'hickade en enda gång' semelfaktiv
b) 'hickade upprepade gånger' process

Ett svenskt verb som potentiellt kan uttrycka semelfaktiv aktion (2a) kan lika gärna uttrycka en process av upprepningar av samma aktion (2b).

Semelfaktiv tolkning kan göras entydig med *till*: *My hickade till*.

Några aktionsartsväxlingar är specifika för små grupper av verb eller enstaka verb, se exempel (3) och (4):

3 a) Prata i ett par minuter!
'håll på med pratandet'
tidsadverbialet bestämmer en process

b) Öppna fönstret i ett par minuter!
'låt fönstret vara öppet'
tidsadverbialet bestämmer det tillstånd som uppstår efter aktionen 'öppna'

Tidsadverbial som *i ett par minuter* brukar bestämma den aktion som verbet direkt anger (3a) men kan i vissa fall istället bestämma det tillstånd som aktionen leder fram till (3b). Den möjliga växlingen gäller en grupp

entydiga tillståndsövergångsverb, t.ex. *lämna*, *stänga*, *sätta sig*, *tända*, *öppna*.

4	a) Bo smakade på kaffet [kortvarig] process	Kaffet smakade surt 'är surt', 'har sur smak', tillstånd
	b) Bo smuttade på kaffet [kortvarig] process	*Kaffet smuttade surt

Verbet *smaka* i (4a) har en växling från vad någon **gör** till hur något **är**; verbet *smutta* i (4b) har inte motsvarande växling. Även om verben i vissa sammanhang kan beskriva samma verklighet, så har de uppenbarligen inte exakt samma lexikala egenskaper.

De flesta verb kan uttrycka mer än en aktionsart, beroende på konstruktion och kontext, och aktionsarten kan också vara oklar i en given kontext. Men alla verb kan inte uttrycka alla mönster, utan de grupperar sig mer eller mindre tydligt. Det finns också en klar tendens i fråga om vilka verbala betydelsetyper som uttrycker vilka aktionsarter, se figur 4.

Verbala betydelsefält	**Typiska aktionsarter**
I. Materiella verb II. Sociala verb	process (*borra*, *pladdra*, *snurra*) gränsrelaterad process (*fylla tanken*, *läsa ut boken*) semelfaktiv (*nysa*, *rycka till*) tillståndsövergång (*anställa ngn*, *få ngt*, *sluta*) [tillstånd: enbart vid generisk/habituell l. metaforisk tolkning]
III. Mentala verb IV. Relationella verb	tillstånd (*veta*, *tro*; *förbli*, *behålla*) tillståndsövergång (*inse*, *upptäcka*; *bli*, *få problem*)

FIGUR 4. Relationen mellan verbala betydelsefält och aktionsarter.

Tillståndsövergångar finns inom alla fyra makrofält, men annars uttrycker materiella och sociala verb typiskt olika sorters processer, varav semelfaktiv kan ses som ett specialfall, medan de mentala och relationella verben normalt bara uttrycker tillstånd och som sagt tillståndsövergångar. Det är en fråga om tendenser, men i grova drag ser det ut så här. Man kan kon-

statera att både verbala makrofält och aktionsart är infallsvinklar som behövs för att beskriva verbens semantik. Notera att det i Hallidays utförliga verbbeskrivning inte ingår någon aktionsartskomponent, men det är alltså fullt möjligt att kombinera verbens makrodomäner med en sådan.

Om man, med stöd av översikten i figur 4, vill slå fast hur svenska verblexem vanligen bidrar till, men inte avgör, aktionsartsbetydelsen i en verbfras, så går det att urskilja tre lexikala grundtyper: tillståndsverb (som är mentala eller relationella), processverb (som är materiella eller sociala) och tillståndsövergångsverb (som kan vara av vilken makrotyp som helst). En motsvarande tredelning i fråga om grundläggande aktionsarter finner man hos t.ex. Durst-Andersen (1992:54–63), Klein (1994, kap. 5), de Swart (1998:351) och Verkuyl (2011:65–67).

5. Lexikal typ på mikronivå

I det här avsnittet ska vi se närmare på enskilda verbs lexikala bidrag. Vi börjar med några tillståndsövergångsverb.

5.1. Tillståndsövergångsverb (TSÖ-verb)

Ett TSÖ-verb anger, när det konstrueras på sitt prototypiska vis, att ett nytt tillstånd inträder, t.ex. *Anna och Pelle gifte sig*; *Hon lämnade Alfa Laval*. Eftersom tillståndsövergången är lexikaliserad i verbet blir det nya tillståndet logiskt implicerat i kontexten. Några exempel ges i (5) och (6):

5 a) Hon öppnade dörren och gästerna gick in
[*öppna* är ett TSÖ-verb]
b) ??Hon nuddade dörren och gästerna gick in

Om vi tänker på vad verbet anger om referensen, så säger *Hon öppnade dörren* att dörren efter själva öppnandet är öppen, medan *Hon nuddade dörren* inte säger något om vad resultatet blir, och (5b) är därför ett exempel som kräver mycket av sin kontext för att vara ett rimligt uttalande.

6 a) Han blev knivhuggen, men överlevde
b) *Han blev mördad, men överlevde
[*mörda* är ett TSÖ-verb]

Samma sak, fast ännu tydligare, gäller för verbet *mörda*. Någon kan bli *illa knivhuggen* och ändå överleva. Ingen kan bli *mördad* och överleva; det nya tillståndet kan inte omedelbart motsägas. Den slutsats vi kan dra är att *öppna* och *nudda* respektive *knivhugga* och *mörda* ger olika semantiska bidrag till sin kontext.

5.2. Domän och profil

Skillnaden mellan TSÖ-verb och andra verb kan beskrivas teoretiskt med Langackers begreppspar *domän* och *profil*, som gäller alla innehållsord, inte bara verb (Langacker 1987:183–189[2], jfr Ekberg 2004:4–5; Nilsson 2019:52–53). En domän är ett erfarenhetsbaserat område, ett utsnitt ur verkligheten som vi har kännedom om och urskiljer kognitivt. En profil är ett ords semantiska bidrag inom en viss domän. Så har ordet *tumme* handen som domän, *ciabatta* har bröd som domän, och *fred* har socialt tillstånd i omvärlden som domän. En domän kan vara ett snävare eller vidare utsnitt ur verkligheten, domäner finns på olika nivåer, och de kan kombineras på olika sätt. Den kognitiva semantiken menar att det inte finns någon gräns mellan omvärldskunskap och språklig kunskap, och relationen mellan domän och profil kan därför vara av många olika slag. Profilen kan syfta på en del av en helhet (som *tumme* och hand) eller ett underbegrepp som *ciabatta* och bröd. Men domänen måste inte vara något som i sig har ett etablerat lexem, utan ibland man får beskriva den så gott man kan, som för *fred*.

Utan anspråk på uttömmande analys kan man således hävda att de materiella verben *knivhugga*, *mörda* och *förolyckas* alla tillhör samma domän, som vi kan kalla ”uppkomst eller åsamkande av allvarlig kroppsskada på animat referent”. Inom denna domän ger verben olika lexikala bidrag. Verbet *knivhugga* bidrar med profilen ’använda kniv för att skada någon’; *mörda* med ’uppsåtligt döda någon’ (SO) och *förolyckas* med ’dö genom olyckshändelse’ (SO). Som vi redan har sett är *mörda* ett TSÖ-verb, medan *knivhugga* inte är det. Och verbet *förolyckas* står just nu och väger mellan den profil som SO anger och profilen ’vara med om allvarlig

2 Notera att Langacker här talar om *base*, inte *domain*, och *profile*. Men eftersom domäner kan utgöra baser och termen domän kan användas om verbbetydelser på många olika nivåer väljer jag att tala direkt om domän och profil. Se Nilsson (2019:52–53) för en termutredning.

olyckshändelse', som är vanlig hos yngre språkbrukare. I det senare fallet har verbet tappat sin lexikala TSÖ-komponent och blivit något annat.

Det finns en likhet mellan den kognitiva semantikens begreppspar domän och profil och klassisk intensionell definition, med vilken man definierar ett ord genom att ange närmaste överbegrepp och därtill lägga minst ett specifikt särdrag, t.ex. när *mala* definieras som 'sönderdela i kvarn' (Svensén 2004:273–276). Det finns också centrala likheter med ramsemantik. Inom ramsemantiken urskiljer man olika typer av ramelement (*frame element*, förkortat FE): "Core FEs are seen to be conceptually necessary for the event depicted by the frame, whereas Non-Core FEs specify more general circumstances such as MANNER, MEANS, PLACE and TIME." (Willich 2022:140.) Core FEs motsvarar valensbundna semantiska roller (jfr nedan) medan non-core FEs snarast motsvarar olika adverbiella betydelser. (Jfr också Fillmore & Baker 2015 och Ruppenhofer et al. 2016:23–25.) Medan ramsemantiken erbjuder en utförlig modell för beskrivning av verbens ramar fokuserar min modell på hur typen av lexikal profil får konsekvenser för verbets samlade semantik.

Langackers uppdelning i domän och profil är också tillämplig på en kontrastiv skillnad som beskrivs av Talmy: Germanska språk har många verb som anger sättet som en rörelse utförs på, men inte rörelsens riktning. Romanska språk har ofta riktningen lexikaliserad i verbet, men inte sättet. (Talmy 2000:60.) Domänen är i samtliga fall "förflyttning", men profilerna är olika. Här några exempel ur den översättningsvetenskapliga litteraturen (Vinay & Darbelnet 1995:51; Lundquist 2007:22):

	Exempel	*Verbets bidrag*
7	a) en. He *swam* across the river	sätt (egen kropp)
	b) fr. Il *traversa* la rivière à la nage	riktning
8	a) sv. Hon *red* in i stallet	sätt (vehikel)
	b) it. *Entrò* a cavallo nella stalla	riktning

På svenska kan vi säga både *Han simmade över floden* (jfr 7a) och *Han korsade floden simmande* (jfr 7b), även om det första är mest naturligt. Poängen är att man kan *korsa floden* hur som helst, så länge man kommer över den. Man kan vada, simma, rida eller ta sig över med båt eller flotte. Både *simma* och *korsa* hör till det materiella fältet, men *simma* talar om

vad subjektsreferenten utför med sin egen kropp. Det gör inte *korsa*. Verbens semantiska bidrag är alltså olika i *simma* och *korsa*. När det gäller *rida in i stallet* är det inte exakt hur kroppen är involverad i aktionen som utgör verbets profil utan vilket transportmedel som behövs – vilket i sin tur styr hur människokroppen kan visualiseras.

5.3. Verb som preciserar en aktants kroppsliga involvering (PKI-verb)

Verb som *knivhugga*, *rida* och *simma* och har någon typ av ”preciserad kroppslig involvering” som lexikal profil. De anger att någon utför något med sin kropp, och detta är lätt att visualisera när verbet används. Verb som *inträda*, *korsa* och *mörda* utförs också av mänskliga kroppar, men verben preciserar inte hur utan istället ett resultat. Verben ensamma ger oss ingen ledtråd om hur själva aktionen gestaltas, även om vi kan ha kännedom om hur vissa aktioner brukar utföras. Några fler exempel, där exempelparen inte är synonyma men ligger inom likartade referentiella domäner:

	PKI-verb	*TSÖ-verb*
9	a) Chefen räckte mig en ros	Jag fick en ros av chefen
	b) Ulla skakade Kalle	Ulla väckte Kalle
	c) Samir bar fram soppan	Samir hämtade soppan
	d) Åke cyklade utom synhåll	Åke försvann bortom kröken
	e) Olle rullade in i rummet	Olle kom in i rummet
	f) Lina skrattade	Lina blev glad

Notera att Ulla kan ’skaka Kalle’ för att väcka honom eller ’väcka honom’ genom att skaka honom, men det ena är inte nödvändigtvis länkat till det andra. Exempelparen illustrerar hur språkbrukaren med sitt val av verb kan konstruera en och samma verklighet på olika sätt, helt enkelt visa sin tolkning av en konkret situation. Vad är då den lexikala skillnaden mellan PKI- och TSÖ-verben i (9)? Alla tillhör makrofältet materiella verb, men relationen mellan subjektsreferenten och själva aktionen är olika. Verben till vänster har lexikala profiler som anger hur subjektsreferentens kropp är involverad i specifika materiella konstellationer. Verben till höger har inte det. Skillnaden kan verka hårfin: Olle är precis lika fysiskt involverad

i aktionen 'kom in i rummet' som i 'rullade in i rummet', ändå visualiserar vi olika saker. För den som känner Olle, och vet att han använder rullstol, är det sak samma, men knappast för den som inte känner Olle. Det handlar om vad verbet uttalar sig om, vad som ligger i den lexikala profilen.

Verb som har den förkroppsligande förmågan kallar jag alltså för PKI-verb, och de finns av olika typer. De allra tydligaste PKI-verben är sättsverb som *fnysa*, *fräsa*, *jogga*, *sparka*, *spotta*, *tugga*, *vinka*. Ibland är som synes till och med kroppsdelen specificerad; man kan inte *fräsa* med foten eller *sparka* med näsan. Instrumentverb som *dammsuga*, *klippa* och *skeda* eller vehikelverb som *cykla*, *rida* och *skida* är också lätta att visualisera som fysiska konstellationer, även om människokroppen då är "förlängd" med något fordon eller verktyg. PKI-verb är alltså inte en enda verbtyp utan åtminstone tre besläktade undergrupper i det materiella makrofältet. (Jfr TSÖ-verben som inte är knutna till någon enskild makrodomän.) Det ska också sägas att PKI-faktorn kan vara olika stark. Verbet *flyga* beskriver vad en kropp gör, men inte så precist; verbet *ryttla* är mycket distinktare.

Observera att termerna TSÖ-verb och PKI-verb pekar ut typer av lexikala profiler, medan den exakta profilen antas vara specifik för varje enskilt verb. TSÖ-verb har profiler som inkluderar en tillståndsövergång (*hämta*, *lämna*); PKI-verb har profiler som inkluderar hur en kropp beter sig rent konkret (*rycka*, *snörvla*). I artikeln fokuserar jag främst på typer av profiler, inte detaljerade profilanalyser.

Ett möjligt belägg för att PKI-verb är den mest grundläggande typen av verb kan man få från barnspråksinlärningen. Svenska barn kan ibland förväxla verben *tända* och *släcka* under lång tid. De säger *tända* när de borde säga *släcka* och vice versa. Båda verben är TSÖ-verb (jfr 10), inte PKI-verb (jfr 11).

10 a) Sam släckte lampan och det blev mörkt
b) ??Sam släckte lampan och det blev ljust
[*släcka ngt* är ett TSÖ-verb]

11 a) Sam tryckte på knappen och det blev mörkt
b) Sam tryckte på knappen och det blev ljust
[*trycka på ngt* är ett PKI-verb]

Exempel (11a) och (11b) är båda logiska, eftersom *trycka på knappen* inte uttalar sig om något resultat, men (10b) är konstig eftersom *släcka* beskriver en övergång till mörker. Min hypotes är att barn utgår ifrån att både *tända* och *släcka* betyder 'trycka på knappen' för att PKI-verb, som har kroppen i centrum, är den första verbtyp de etablerar. För barnet är det tryckandet på knappen som verbet syftar på, det man **gör**, även om det är tillståndsförändringen som är det roliga med leken.

När man väl har urskilt PKI-faktorn, kan man gå till andra delar av verbbeståndet och upptäcka att faktorn har beskrivningspotential på fler ställen, se exempel (12).

	PKI-verb	*inte PKI-verb, inte TSÖ-verb*
12	a) Hunden tuggade på bollen	Hunden lekte med bollen
	b) Barnen slogs	Barnen bråkade
	c) Åke hamrar på tangenterna	Åke jobbar
	d) Åsa dammsuger	Åsa städar
	e) Zeyna hackade grönsaker	Zeyna lagade en god soppa
	f) Pelle kavlade degen	Pelle bearbetade degen

Bland verben i den vänstra spalten har vi olika PKI-verb. Bland verben i den högra spalten finner vi både aktivitetsverb och produktionsverb. Det finns alltså många mänskliga och materiella verb, för aktioner som människor utför med sina kroppar, som inte är PKI-verb och inte heller TSÖ-verb. Det är en grupp praktiskt halvinzoomade verb. Människor eller djur utför dem fysiskt, men exakt **hur** anges inte av verbet. Ett, av många, sätt att jobba är att hamra på tangenter, ett moment av att laga soppa kan vara att hacka grönsaker, och det kan ingå i hundens lek att tugga på en boll etc. Genom sitt val av verb kan språkbrukaren konstruera en verklig situation i aktioner på mer än ett sätt.

Bland verblexemen, såväl PKI-verb som andra, kan graden av precisering variera, och preciseringen kan vara lexikaliserad i verbet eller tillhandahållas av element i verbfrasen. Några exempel ges i (13) och (14).

PKI-verb

	mindre specifikt	*mer specifikt*
13	a) simma	crawla
	b) gå	släntra
	c) åka	cykla
	d) dansa	dansa vals/tango

aktivitets- och produktionsverb

	mindre specifikt	*mer specifikt*
14	a) arbeta	extraknäcka
	b) spela	spela fiol/golf
	c) bygga	bygga ett hus/ett sandslott
	d) baka	baka kakor/potatis

Simma är ett PKI-verb men *crawla* är mer specifikt, och det enda verb vi har för ett enskilt simsätt. (Jfr hästverb som *galoppera*, *skritta*, *trava* och *tölta*.) Att *gå ner till stranden* är inte lika gestaltande som att *släntra ner till stranden*. Verb som *åka* och *köra* är halvpreciserade, de kräver ett fordon, men anger inte vilket; verbet *cykla* är helpreciserat. *Dansa* är ett PKI-verb, men vi kan specificera mera exakt vilket dansmönster som följs. Varken *arbeta* eller *extraknäcka* indikerar vilken sorts jobb som utförs, ändå är *extraknäcka* ett snävare verb. Aktivitetsverb som *spela* och *leka* är inte PKI-verb, men de kan ändå göras mer preciserade av andra led i verbfrasen: att 'spela fiol' eller 'spela golf' är två kroppsligt mycket olikartade sysselsättningar. I ett fall som *laga mat* respektive *laga en cykel* är skillnaden så pass stor att vi nog uppfattar verbet *laga* som polysemt. Men även här gäller att verbets bestämningar preciserar betydelsen utan att förvandla själva verbet till ett PKI-verb.

En viktig del av domän-profil-modellen är att de delar som inte uttrycks av verbets profil är närvarande ändå, för att de ligger i själva domänen. Om ett verb är ett materiellt verb, oavsett vilken profil det har, så är det mycket som följer med på köpet, och det som ligger i domänen kan lyftas fram vid behov genom olika typer av preciseringar i verbfrasen. Alla aktioner som konstrueras som materiella äger rum på en viss plats, vid en viss tid och har en viss tidsutsträckning, även om den är kortvarig. Är rörelse involverad så aktualiseras omedelbart vägschemat med källa och mål. Olika faser av en aktion kan också fokuseras, t.ex. början, pågåen-

det, slutet. Detta gör Langackers begreppspar domän och profil särskilt förklarande i fråga om verbsemantik. Egenskaperna finns där och är möjliga att lyfta fram även om de inte ingår i profilen. Några exempel är:

15	Kalle sprang	i morse	[tid]
		i tio minuter	[durativitet]
		i riktning mot kyrkan	[mål]
		iväg	[källa]
16	Kalle snyftade	i köket	[plats]
		högljutt	[preciserat sätt]
		med vilje	[agentivitet]
		fortfarande	[pågående]

Min uppfattning är att själva materialiteten, den fysiska konstellationen, utgör profil endast hos PKI-verb. Verbets lexikaliserade bidrag är då att ange sätt, instrument eller vehikel, och dessa betydelseelement expliciterar aktionens materiella egenskaper, t.ex. *flaxa*, *pensla* och *segla*. Hos materiella verb som inte är PKI-verb, t.ex. *resa* och *skriva*, ligger materialiteten istället i domänen. På motsvarande sätt är det ytterst få svenska verb som lexikaliserar durativitet (*fortgick*, *höll på*, *varade*), medan någon form av durativitet ligger i domänen för de flesta verb. När det gäller pro-verb som *hända* och *göra*, som kan utpeka ospecificerade händelser eller aktiviteter, är det rimligt att tro att de lexikaliserar drag (t.ex. dynamiskhet eller aktivitet) som hos andra verb inte ingår i profilen utan finns i domänen. Min version av domän-profil-modellen postulerar således att en typ av semantisk komponent som ingår i domänen för vissa verb utgör profil för andra. På så sätt finns det ingen skarp gräns mellan encyklopedisk och språklig kunskap, men enligt min uppfattning är själva förekomsten av lexikala profiler ett lingvistiskt fenomen, inte ett allmänt kognitivt.

Med en verbmodell som denna blir relationen mellan verbets lexikala aktionsartsbidrag (tillstånd, process eller tillståndsövergång) och de fem aktionsartsmönster som brukar urskiljas (jfr figur 3) en fråga om vilken typ av drag i domänen som språkbrukaren faktiskt vill explicitera genom sin konstruktion av verbfrasen: *vackla* är lexikalt ett PKI-verb (och därmed ett processverb); *vackla runt* är aktionsartsmässigt en process, medan

vackla hem är en gränsrelaterad process. Verbets lexikala profil är således mer grundläggande än dess aktionsart.

Det ska också sägas att gränsen mellan domän och profil generellt är variabel. Ett verb har en viss profil, men det finns alltid också en domän som innefattar saker som kan bli mer kommunikativt intressanta, och då kan profilen förskjutas mer eller mindre så att polysemi uppstår (jfr Ekberg 2004).

5.4. Relationella verb

Nu lämnar vi det materiella fältet för en snabb titt på några relationella verb. Relationella verb är entydiga: de anger aldrig preciserad kroppslig involvering.

	Relationella verb	*Jfr PKI-verb*
17	a) Kim är/blev förkyld	Kim snörvlar och hostar
	b) Kim har/fick influensa	
	c) Pojken heter Kim	Mamma ropar på Kim

Vid relationella verb anger själva verbet ingenting om hur kroppen förhåller sig. Den väsentliga informationen kommer i predikativet eller objektet. Att 'vara förkyld' eller 'ha influensa' kan involvera en mängd fysiskt konkreta obehag, men själva verbet utpekar inte dessa. Med ett verb som *heta* blir det tydligt att aktionen är fullständigt lösriven från vad någon fysiskt ägnar sig åt. Pojken 'heter Kim' oavsett om han sover eller är vaken, är närvarande eller inte. Att 'heta Kim' är bara att förknippas med en etikett, inte att vara kroppsligt involverad i någon aktion.

En tydlig skillnad mellan relationella och materiella verb gäller kombinerbarhet och samtidighet. Vid relationella verb är det logiken som avgör hur många predikat som samtidigt kan kopplas till en aktant. Det är fullt möjligt att samtidigt hävda att *Kajsa var gift / blev glad / hade en bil / ägde en båt / egentligen hette Katarina* etc. I fråga om relationella predikat kan man tillskriva en och samma referent hur många samtidiga attribut som helst, så länge de inte krockar logiskt. Vid PKI-verb är det kroppsliga begränsningar som avgör vad som kan gälla samtidigt, se exempel (18) och (19).

18 a) Karin gråter och skrattar samtidigt
b) Karl cyklar och sjunger samtidigt
c) Kaj knådar och nyser samtidigt

19 a) *Karim applåderar och vinkar samtidigt
b) *Kalle gäspar och visslar samtidigt
c) *Katrin sitter och står samtidigt

Med materiella verb kan man inte kombinera aktioner om det är fysiskt orimligt eftersom en viss kroppsdel bara kan ingå i en preciserad aktion i taget. Man kan 'knåda' och 'nysa' samtidigt – även om det är otrevligt. Däremot kan man inte 'sitta' och 'stå' samtidigt, vilket för oss fram till de viktiga svenska verben *ligga*, *sitta* och *stå*.

5.5. Posityrverb

Verben *ligga*, *sitta* och *stå* är speciella i svenskan på flera sätt. Vi använder dem bland annat för att tala om var inanimata referenter befinner sig i mycket högre grad än vad man gör i icke-nordiska språk – ofta i förening med semantiskt finlir.

20 a)	Boken ligger på hyllan	Boken står i hyllan
b)	Lampan står på golvet	Lampan sitter i taket
c)	Steken ligger på bordet	Steken står på bordet

Det krävs ganska mycket analys för att förklara varför en lampa 'sitter i taket' och en stek som 'ligger på ett bord' förmodligen är rå, medan ett stek som 'står på ett bord' sannolikt är tillagad.

Vi använder också verben *ligga*, *sitta* och *stå* i pseudosamordning (jfr Blensenius 2015: delstudie III) och presenteringskonstruktion (jfr Thyberg 2020, kap. 8):

21 a) Hon låg och läste i soffan
b) Levi satt och spelade DotA halva natten

22 a) Det sitter en katt på trappan
b) Det står en polisbil på gatan

Vill man förstå svenskans verbsystem, bör man försöka förstå dessa verb. För en aktionsartsforskare är *ligga*, *sitta* och *stå* svåranalyserade eftersom de ligger precis på gränsen mellan tillstånd och processer, på det ställe i aktionsartssystemet där en av de viktigaste skiljelinjerna går, den mellan dynamiska och icke-dynamiska aktioner (för en diskussion, se Maienborn 2008). En dynamisk aktion innebär alltid att något händer, sker eller förändras. En icke-dynamisk (eller statisk) aktion innebär ingen förändring, se figur 5.

Aktionsart	**Exempel**	**Dynamiskt eller inte**
Tillstånd	*vara glad, äga en cykel*	Icke-dynamiskt
Process	*regna, simma runt*	Dynamiskt
Gränsrelaterad process	*regna bort, simma till stranden*	Dynamiskt
Tillståndsövergång	*avgå, dö, upptäcka*	Dynamiskt
Semelfaktiv	*hicka, hoppa till, hosta*	Dynamiskt

FIGUR 5. Dynamiska och icke-dynamiska aktionsarter.

Tillstånd är icke-dynamiska medan alla andra aktionsarter är dynamiska; de inbegriper en förändring – om än aldrig så liten. Att ’flämta till’ eller ’bli trött’ räcker alltså för att en aktion ska klassas som dynamisk. Verben *ligga*, *sitta* och *stå* reagerar på de flesta aktionsartstest som om de vore tillståndsverb; inget händer, det bara är eller föreligger, icke-dynamiskt (jfr Christensen 1995). Men man kan inte ’sitta’ och ’stå’ samtidigt, för att det är fysiskt omöjligt, och detta är en renodlad PKI-egenskap. Och det är bara materiella verb som kan ha preciserad kroppslig involvering som lexikal profil. Slutsatsen måste vara att *ligga*, *sitta* och *stå* är statiska och PKI-verb samtidigt.

Verben *ligga*, *sitta* och *stå* kallas ibland för positionsverb eftersom de anger var något finns: *Boken ligger på hyllan*; *Bo sitter i köket*; *Vinet står i kylen*. Men deras positionsbestämmande kraft är sekundär. Materiella aktioner måste alltid utspela sig på en viss plats, det är en egenskap som följer med på köpet, och det gäller ’strö va i skogen’ och ’shoppa på stan’ i lika hög grad som ’ligga på soffan’. Lokaliserbarhet ingår i alla materiella verbs domäner. Verben *ligga*, *sitta* och *stå* har som lexikal profil att ange kroppsställning, och därför kan de inte beteckna samtidiga aktioner hos konkreta referenter. Jag föreslår att vi, i analogi med engelskans *posture*

verbs (jfr t.ex. Newman 2002), kallar dem posityrverb på svenska.[3] Dessa verb är semantiskt egenartade eftersom de är både statiska och PKI-verb. En sådan semantisk struktur kan inte beskrivas inom de vanliga aktionsartsmodellerna eftersom dessa inte frågar efter verbens lexikala profiler.

6. Semantiska roller

Till sist något om verbens semantiska roller. De led som ingår i verbets syntaktiska valens har semantiska roller i relation till verbet och övriga led som ingår i valensen. Semantiska roller har liksom aktionsarter en oklar ontologisk status, men de flesta forskare är eniga om att vi med rollernas hjälp kan förklara vissa systematiska språkliga mönster. Rollerna är både knutna till vissa syntaktiska typer och gränsöverskridande mellan syntaktiska typer, och det är därför vi behöver dem. I exempel (23) till (26) åsyftar AGENS en referent som avsiktligt och medvetet utför en aktion, MOTTAGARE den som tjänar på en aktion och FÖREMÅL den eller det som blir utsatt för aktionen.

23	Liv	ger	katten	mjölk
	AGENS		MOTTAGARE	FÖREMÅL
24	Liv	ger	mörten	till katten
	AGENS		FÖREMÅL	MOTTAGARE
25	Ulf	läste	boken	
	AGENS		FÖREMÅL	
26	Ulf	läste	i boken	
	AGENS		FÖREMÅL	

Som synes kan rollen MOTTAGARE uttryckas både med ett indirekt objekt (*katten*) och med ett adverbial (*till katten*). Och även om det är en aktionsartsskillnad mellan att *läsa boken* och att *läsa **i** boken*, så är det ingen skillnad i rollfördelningen. Det är Ulf som läser och boken som blir läst.

3 Jfr SAOB: POSITYR (bet. 1): ”ställning l. attityd som person l. djur intar; (kropps)ställning, hållning”.

Liksom i fråga om aktionsartsmodeller finns det många modeller över semantiska roller. De som har vägts in här är följande: SAG (kap. 7 (Verb), §4–§10); Holmberg & Karlsson (2006, kap. 4); Platzack (2010:71–76); Halliday & Matthiessen (2014, kap. 5); Saeed (2016:150–155). Notera att man inom SFG inte separerar makrofälten och rollerna utan beskriver båda aspekterna samlat inom den så kallade transitivitetsanalysen (för en översikt, se Halliday & Matthiessen 2014:219). Med svenska termer sägs till exempel att ”mentala processer” har deltagarna ”upplevare” och ”fenomen” och att ”verbala processer” har deltagare som ”talare” och ”lyssnare” (Holmberg & Karlsson 2006:102). Trots att de teoretiska infallsvinklarna och de enskilda termerna skiljer sig åt, är det inte så svårt att skissera en modell över semantiska roller som utgör en minsta gemensam nämnare mellan de anförda modellerna och samtidigt synliggör de fyra makrofälten. En sådan modell skulle kunna se ut som i Figur 6:

1. Roller i de materiella och sociala fälten	KROPP, AGENS, ORSAK; FÖREMÅL, MOTTAGARE; TEMA, PATIENT, RESULTAT; PLATS, TID, INSTRUMENT
2. Roller i det mentala fältet	UPPLEVARE, FENOMEN
3. Roller i det relationella fältet	BÄRARE

FIGUR 6. Relationen mellan verbala betydelsefält och semantiska roller.

Flest roller ser vi i det sammanslagna materiella och sociala fältet, vilket indikerar att verbsemantikens kärna befinner sig här. Verbens mest prototypiska uppdrag är att urskilja konstellationer där människan agerar som kropp i den fysiska världen, med eller utan social kontext, och därför har språket många olika roller och rollkonstellationer här. Mentala och relationella verb har ett betydligt smalare uppdrag, och därför behövs färre roller där.

Här är inte platsen att fördjupa rollsemantiken, men ett teoretiskt problem som inte har en given lösning är hur man ska betrakta referenter som gestaltar aktioner oavsiktligt. Med rollen AGENS brukar man, som sagt, åsyfta en referent som avsiktligt och medvetet utför en handling. Men namnet på referenter som gestaltar en aktion oavsiktligt varierar. SAG kallar den rollen för FÖREMÅL (t.ex. *Solen försvann*), vilket också är den

typiska objektsrollen (t.ex. *B slår A*; SAG bd 2:508). Den systemisk-funktionella grammatiken kallar den AKTÖR och MEDIUM (t.ex. *Snön faller*; Holmberg & Karlsson 2006:110). Saeed kallar den THEME (t.ex. *The car rose*; 2016:154), och Platzack försvenskar till TEMA (2010:74). Enligt min mening skjuter dessa termer förbi målet, när det i grunden handlar om vilken fysisk kropp, biologisk eller ej, som gestaltar en aktion. Materiella aktioner utförs alltid av fysiska kroppar – det ligger i själva definitionen av en materiell aktion. Jag föreslår därför att rollen kallas KROPP, eftersom denna etikett har ett större förklaringsvärde.

Om man urskiljer rollen KROPP och håller isär den från AGENS skulle några enkla svenska satser få följande analys, där rollen KROPP kan definieras som ”den eller det som materiellt gestaltar den av verbet beskrivna aktionen”.

27 a) Kaj gråter
b) Kaj sover
KROPP

28 a) Kaj springer
b) Kaj gräver en grop
KROPP + AGENS

Vanliga mänskliga aktiviteter kräver att man använder sin kropp för att utföra dem samtidigt som många av dem görs avsiktligt och kontrollerat, alltså agentivt. Skillnaden mellan *Kaj gråter* och *Kaj gräver* är alltså en skillnad som har med kontroll att göra, inte kroppslighet. Därför föreslår jag en analys med dubbla roller när en kroppslig aktion är agentiv.[4] Analysen i (27) fungerar lika bra för inanimata materiella referenter, t.ex. *Bäcken porlar* eller *Båten sjönk*, men då är förstås rollen AGENS inte aktuell.

En fördel med en analys som uttryckligen använder rollen KROPP för den referent som gestaltar verbets aktion materiellt är att man enkelt kan beskriva skillnaden mellan ett vanligt transitivt satsmönster och ett kausativt mönster:

4 Jfr SAG (bd 2:506): ”Ett agentivt verb kan ha ett enda aktantled, som ofta samtidigt anger agens och aktionens föremål: *A springer*.” (Originalets kursiv.)

	vanligt transitivt mönster	*jfr*
29	a) Elsa hämtade smöret	*Smöret hämtade
	b) Elsa pumpade bollen	*Bollen pumpade
	c) Elsa byggde sandslottet	*Sandslottet byggde

	kausativt mönster	*jfr*
30	a) Elsa smälte smöret	Smöret smälte
	b) Elsa rullade bollen	Bollen rullade
	c) Elsa välte tornet	Tornet välte

I både (29) och (30) är Elsa fysiskt involverad i en materiell aktion, hon **gör** något. I (29) är det Elsas kropp som hämtar, pumpar och bygger, men i (30) är det inte Elsa som smälter, rullar eller välter. Det gör nämligen objektsreferenterna (smöret, bollen eller tornet). Och det är bara hos verben i (30) som vi kan se konstruktionsväxling mellan transitiv och intransitiv sats och bara dessa som kan omskrivas med det kausativa verbet *få*: *Elsa fick smöret att smälta*. *Elsa fick bollen att rulla*, men inte **Elsa fick smöret att hämta* eller **Elsa fick sandslottet att bygga*. Eftersom det förhåller sig så, föreslår jag följande analys, där KROPP alltså definieras som ”den eller det som materiellt gestaltar den av verbet beskrivna aktionen”, AGENS som ”den som avsiktligt och medvetet utför en aktion” och FÖREMÅL som ”den eller det som utsätts för den av verbet beskrivna aktionen”:

	vanligt transitivt mönster		
31	Elsa	pumpade	bollen
	KROPP+AGENS		FÖREMÅL

	kausativt mönster		
32	Elsa	rullade	bollen
	AGENS		KROPP+FÖREMÅL

Det är viktigt att inte glömma att KROPP betyder ’förkroppsligare av verbaktionen’. Elsa har fortsatt en kropp när hon rullar bollen, men det är inte hennes kropp som rullar. En intressant detalj är att den aktant som aktualiseras av verb som *jäsa*, *krympa*, *koka*, *landa*, *rulla*, *smälta*, *snurra*, *tina* och *välta* (uttryckt som objekt eller subjekt) troligen måste gestalta aktionen med preciserad kroppslig involvering. Vid konstruktionsväx-

lande verb som dessa krävs det alltså att verbet beskriver vad en KROPP gestaltar materiellt.

Liksom ifråga om aktionsartsväxling varierar möjligheterna till konstruktionsväxling. *Han skruvade bollen i mål* kan till exempel inte skrivas om till **Bollen skruvade i mål*. Och även om *Potatisen kokar* har motsvarigheten *Eva kokar potatis*, så saknar *Potatisen svalnar* motsvarigheten **Eva svalnar potatisen*. Och ett fall som *Bo vände båten* och *Båten vände* tolkar vi kanske snarare som en metonymi (BÅTEN FÖR DEN SOM STYR BÅTEN). Även om en grupp av verb uppvisar ett specifikt syntaktiskt mönster, måste inte semantiskt likartade verb ansluta till det. Varje utförlig beskrivning av verb måste ta hänsyn till deras individuella egenskaper.

Om vi nu zoomar ut från detaljerna och tänker bredare, så kan man konstatera att svenska ordböcker inte använder rolletiketter i sina konstruktionsbeskrivningar. Man klarar sig med *ngn* för animata aktanter, *ngt* för inanimata och *ngn/ngt* när distinktionen inte behövs. Hur kan det komma sig? Min tanke är att det beror på att det mesta av rollsemantiken ligger i domänerna, inte i profilerna. Om ett verb är ordentligt beskrivet, så att ordboksbrukaren förstår både vilka vidare och snävare domäner det tillhör, vilken lexikal profil det har och vilka konstruktioner det kan ingå i, då kan man förmoda att de mer specifika rollerna faller ut av sig själv, för att de tillhör domäner som vi redan har förståelse för. Vid ett mentalt verb blir den animata aktanten UPPLEVARE och det som upplevs blir ett FENOMEN, vid ett sägeverb blir den som säger något TALARE och en annan person kan bli LYSSNARE och det sagda blir UTSAGA (jfr Holmberg & Karlsson 2006:102), och vid ett typiskt mänskligt handlingsverb står *ngn* för KROPP/AGENT och *ngn/ngt* för FÖREMÅL etc. Även om vi behöver mer exakta rolletiketter för att förklara syntaktiska mönster så verkar vi inte behöva dem för att beskriva verblexem.

7. Avslutning

I denna artikel har jag försökt tänka ihop verbala makrofält med aktionsart och semantiska roller. Och vägen fram har varit att för enskilda verb urskilja domän och profil i den kognitiva semantikens bemärkelse. Med detta synsätt hamnar stora delar av aktionsarts- och rollsemantiken i domänerna, som potentiella betydelseelement som språkbrukarna genom sin omvärldskunskap har kännedom om. Och när verben används i satser och meningar

kan språkbrukaren välja att lyfta upp och explicitera inslag i domänen och uttrycka dem som led i verbfrasen eller låta bli. Det utmärkande för min modell är att jag konsekvent försöker avgöra de enskilda verbens typ av lexikal profil (TSÖ-verb, PKI-verb eller något annat), eftersom det i hög grad är denna som förklarar verbets övriga semantiska egenskaper.

Som sagt i inledningen är denna modell över svenska verbs lexikala semantik inte avsedd att fungera som manual vid lexikografiskt arbete. Min förhoppning är ändå att universitetssemantiken har något att erbjuda ordbokssemantiken i fråga om hur man överordnat kan tänka kring verb. En sak är i alla fall säker. Jag har haft stor nytta av ordböckernas beskrivningar av verb under alla de år som jag arbetat teoretiskt med verbsemantik (jfr Christensen 2002).

Litteraturförteckning

Blensenius, Kristian 2015. *Progressive constructions in Swedish*. (Göteborgsstudier i nordisk språkvetenskap 25.) Göteborg: Göteborgs universitet.

Christensen, Lisa 1995. *Svenskans aktionsarter*. (Nordlund 20.) Lund: Lunds universitet.

Christensen, Lisa 2002. Universitetssemantik och ordbokssemantik. I: Mattisson, Anki, Per Stille, Gunilla Swietlicki & Bo-A. Wendt (red.), *Alla ord är lika roliga. Festskrift till Lars Svensson 28 februari 2002*. Stockholm: Svenska Akademien, 24–39.

Christensen, Lisa 2010. *Early Verbs in Child Swedish – a Diary Study on two boys*. (Nordlund 30.) Lund: Lunds universitet.

Croft, William 2012. *Verbs: Aspect and Clausal Structure*. Oxford Scholarship Online.

Durst-Andersen, Per 1992. *Mental Grammar. Russian Aspect and Related Issues*. Columbus, Ohio: Slavica Publishers.

Ekberg, Lena 1989. Gå till anfall *och* falla i sömn. *En strukturell och funktionell beskrivning av abstrakta övergångsfraser.* (Lundastudier i nordisk språkvetenskap A 43.) Lund: Lund University Press.

Ekberg, Lena 2004. Grundbetydelse och förändringsprinciper hos relationella lexem. Exemplet *följa*. I: *Tre uppsatser om semantisk förändring hos relationella lexem*. (Nordlund 24.) Lund: Lunds universitet, 1–24.

Fillmore, Charles J. & Collin Baker 2015 [2010]. A Frames Approach to Semantic Analysis. I: Heine, Bernd & Heiko Narrog (red.), *The Oxford Handbook of Linguistic Analysis*. 2 uppl. Oxford: Oxford University Press, 791–816.

Halliday, M.A.K. & Christian M.I.M. Matthiessen 2014. *Halliday's Introduction to Functional grammar*. 4 uppl. London & New York: Routledge.

Holm, Lisa (u.å.). *Vendler revisited – from a Swedish point of departure*.

Holmberg, Per & Anna-Malin Karlsson 2006. *Grammatik med betydelse. En introduktion till funktionell grammatik*. (Ord och stil 37.) Uppsala: Hallgren & Fallgren.

Jakobsson, Ulrika 1996. *Familjelika betydelser hos* stå, sitta *och* ligga. *En analys ur den kognitiva semantikens perspektiv*. (Nordlund 21.) Lund: Lunds universitet.

Kinn, Torodd, Kristian Blensenius & Peter Andersson 2018. Posture, location, and activity in Mainland Scandinavian pseudocoordinations. *CogniTextes* 18. <https://journals.openedition.org/cognitextes/1158>. Hämtat mars 2023.

Klein, Wolfgang 1994. *Time in Language*. London & New York: Routledge.

Langacker, Ronald W. 1987. *Foundations of Cognitive Grammar*. Vol. I. Stanford, California: Stanford University Press.

Lundquist, Lita 2007. *Oversættelse*. Frederiksberg: Forlaget Samfundslitteratur.

Maienborn, Claudia 2008. On Davidsonian and Kimian States, I: Comorovski, Ileana & Klaus von Heusinger (red.), *Existence: Semantics and Syntax*. (Studies in Linguistics and Philosophy 84.) Dordrecht: Springer, 107–130.

Newman, John 2002. A cross-linguistic overview of the posture verbs 'sit', 'stand', and 'lie'. I: Newman, John (red.), *The Linguistics of Sitting, Standing and Lying*. (Typological studies in language 51.) Amsterdam/Philadelphia: John Benjamins Publishing Company, 1–24.

Nilsson, Pär 2019. *Bildliga betydelser i SAOB. Om beskrivningen av betydelseutvecklingsmekanismer analyserad ur ett kognitivt semantiskt perspektiv.* (Lundastudier i nordisk språkvetenskap A 79.) Lund: Lunds universitet.

Platzack, Christer 2010. *Den fantastiska grammatiken. En minimalistisk beskrivning av svenskan*. Stockholm: Norstedts.

Ruppenhofer, Josef, Michael Ellsworth, Miriam R. L. Petruck, Christopher R. Johnson, Collin F. Baker & Jan Scheffczyk 2016. *FrameNet II: Extended Theory and Practice*. <https://framenet.icsi.berkeley.edu/fndrupal/index.php?q=the_book>. Hämtat mars 2023.

Saeed, John I. 2016 [1997]. *Semantics*. 4 uppl. Chichester: Wiley Blackwell.

SAG = Teleman, Ulf, Staffan Hellberg & Erik Andersson 1999. *Svenska Akademiens grammatik*. Stockholm: Svenska Akademien & Norstedts Ordbok.

SAOB = *Ordbok över svenska språket utgiven av Svenska Akademien* 1898–. Lund: Gleerups.

Sasse, Hans-Jürgen 2002. Recent activity in the theory of aspect: Accomplishments, achievements, or just non-progressive state? *Linguistic Typology* 6:2, 199–271.

Smith, Carlota S. 1997 [1991]. *The Parameter of Aspect*. 2 uppl. Dordrecht/Boston/London: Kluwer Academic Publishers.

SO = *Svensk ordbok utgiven av Svenska Akademien* 2021 [2009]. 2 uppl. Stockholm: Svenska Akademien. <svenska.se>. Hämtat mars 2023.

de Swart Henriëtte 1998. Aspect Shift and Coercion. *Natural Language and Linguistic Theory* 16:2, 347–385.

Svensén, Bo 2004 [1987]. *Handbok i lexikografi. Ordböcker och ordboksarbete i teori och praktik*. 2 uppl. Stockholm: Norstedts Akademiska Förlag.

Talmy, Leonard 2000. *Toward a Cognitive Semantics*. Vol. II. Cambridge, Massachusetts; London, England: The MIT Press.

Thyberg, Kajsa 2020. Det-*konstruktioner i bruk*. *En systemisk-funktionell analys av satser med icke-referentiellt* det *i modern svenska*. (Göteborgsstudier i nordisk språkvetenskap 41.) Göteborg: Göteborgs universitet.

Vendler, Zeno 1957. Verbs and times. *The Philosophical Review* 66:2, 143–160.

Verkuyl, Henk J. 2011 [1993]. *A Theory of Aspectuality. The Interaction between Temporal and Atemporal Structure*. Online publ. Cambridge University Press.

Viberg, Åke 2008. Swedish verbs of perception from a typological and contrastive perspective. I: de los Ángeles Gómez Gonzáles, Maria, J. Lachlan Mackenzie & Elsa M. González Álvarez (red.), *Languages and Cultures in Contrast and Comparison*. (Pragmatics & Beyond New Series 175.) Amsterdam/Philadelphia: John Benjamins Publishing Company, 123–172.

Vinay, Jean-Paul & Jean Darbelnet 1995 [1958]. *Comparative Stylistics of French and English. A Methodology for Translation.* (Translated and edited by Juan C. Sager & M.-J. Hamel.) Amsterdam/Philadelphia: John Benjamins Publishing Company.

Willich, Alexander 2022. Introducing Construction Semantics (CxS): a frame-semantic extension of Construction Grammar and constructicography. *Linguistics Vanguard* 8:1, 139–149.

Sektionsföredrag och bidrag utifrån posterpresentationer

Mot en harmonisk lemma-lexemmodell och ordklassuppsättning

Kristian Blensenius

The so-called lemma-lexeme model is a lexical principle applied in The Contemporary Dictionary of the Swedish Academy, SO, and The Swedish Academy glossary, SAOL, for determining the character of lexical entries. The principle distinguishes formal properties, e.g., inflection, of the lexical entry (lemma) and content, i.e., meaning (lexeme). This article discusses some cases of different implementation of the framework in SO and SAOL, leading to a discussion of possible harmonization of the dictionaries regarding what lexical principles and set of word-classes should be used.

NYCKELORD: lemma-lexemmodellen, harmonisering, ordböcker, morfologi, ordklassbestämning

1. Inledning[1]

I inledningen till *Svensk ordbok utgiven av Svenska Akademien*, SO, står det att den så kallade *lemma-lexemmodellen* tillämpas i såväl *Svenska Akademiens ordlista* (SAOL) som SO 2009 (SO 2009:IX; se även SAOL 14:XVIII och Malmgren 2014:86). Denna teoretiska modell (Allén 1999 [1981]) innebär bland annat att ord som överensstämmer med avseende på form behandlas som *ett* uppslagsord i ordboken. Syftet med denna artikel är att undersöka hur SO och SAOL genomför lemma-lexemmodellen, med fokus på några skillnader mellan ordböckerna och följderna för ordklassuppsättningarna i ordböckerna.

1.1. Lemma-lexemmodellen i SO och SAOL

SO och SAOL är enspråkiga ordböcker för svenskt samtidsspråk. SO har en deskriptiv inriktning medan SAOL har en mer normativ, och båda innehåller information om ordklass och böjning. SAOL är mer inriktad

1 Tack till granskarna för värdefulla kommentarer.

på produktion, med fokus på stavning och böjning, medan SO fokuserar på betydelsebeskrivningar och information om hur ord används. SAOL ger information om samtliga brukliga böjningsformer, medan SO ger ett urval av former.

SO innehåller ca 65 000 uppslagsord, medan SAOL innehåller ca 125 000. Ett uppslagsord eller, snarare, en grupp likstavade ordformer inom samma ordklass, ett *lemma*, etableras principiellt i båda ordböckerna för ord som överensstämmer formellt, framför allt avseende stavning, ordklass och böjning, men även avseende uttal och segmentering[2] (men inte t.ex. avseende syntaktisk valens; Borin 2008:64). Vid homografi markeras lemmat med framförställd indexsiffra, t.ex. ²**bok** (fetstil indikerar uppslagsord). Det innehåll som lemmat uttrycker visar sig som huvudbetydelser av lemmat, *lexem*, som indikeras med siffra efter uppslagsordet, t.ex. **krona 2**. Huvudbetydelserna delas i tillämpliga fall också in i underbetydelser.

I praktiken etableras lemmastatus enligt ett schema (se Allén 1999 [1981]:275)[3]. Två ords grundformer jämförs, t.ex. *fil* och *bil*. Eftersom dessa former skiljer sig åt med avseende på stavning, ska de vara olika lemman. Om vi i stället jämför ord med tillhörande former (*flektionsserier*), som *fil* (best. sing. *filen* plur. *filar*) och *fil* (*filen filer*), ser vi att *fil* och *fil* passerar stavningskriteriet, de stavas likadant, men de divergerar i plural. Vidare går det inte att fritt byta *filar* mot *filer* utan att huvudbetydelsen förloras. Således ska de utgöra olika lemman. Exempel som *nit* 'lott utan vinst' (*niten nitar/niter*) uppvisar förhållandevis fri variation i plural men uppvisar samtidigt opposition i förhållande till *nit* 'bult' (*niten nitar*), eftersom pluralformen *niter* inte kan användas i betydelsen 'bult'. Därmed är *nit* 'bult' och *nit* 'lott utan vinst' att betrakta som olika lemman. Två flektionsserier som dock förs till ett och samma lemma är *svan* (*svanen svanar*) och *svan* (*svanen svanor*), där pluralformen *svanar* och den ålderdomliga *svanor* i princip är utbytbara utan avsevärd semantisk förändring i ett visst textsammanhang.

2 Segmenterings- och uttalsskillnader framträder t.ex. hos substantiven **bil-drulle** och **bild-rulle**, **förband** och **för-band** med flera (bindestrecket anger sammansättningsgräns). Dessa hanteras som olika lemman i SAOL, trots att de böjs på samma sätt. I SO ses en liknande lemmauppdelning hos **förtrycka** resp. **för-trycka**.

3 Se Ralph, Järborg & Allén (1977:56–57) och Järborg (1989:5–7) för en genomgång av lemma-lexemmodellens tillämpning i Lexikalisk databas, som SO bygger på.

Etymologi är inte lemmaskiljande, och detta medför att formellt lika ord kan föras till ett och samma lemma trots olika ursprung: i såväl SAOL som SO är t.ex. både *ren* ’djur’ och *ren* ’kant, markremsa’ huvudbetydelser under uppslagsordet ²**ren**, inte på grund av etymologiskt släktskap, utan på grund av att de båda orden bl.a. delar ordklass, stavning och böjning. Se figur 1.

FIGUR 1. ²**ren** 1 i SAOL och SO på Svenska.se.

Ett verb som *kröka* behandlas på ett annat sätt. Det delas, trots etymologisk relation, upp på två lemman, ¹**kröka** (’kröka på armen, dvs. supa’) och ²**kröka**, vilket kan tillskrivas dess olika böjningsmönster: ¹**kröka** (*krökade krökat*) och ²**kröka** (*krökte krökt*).

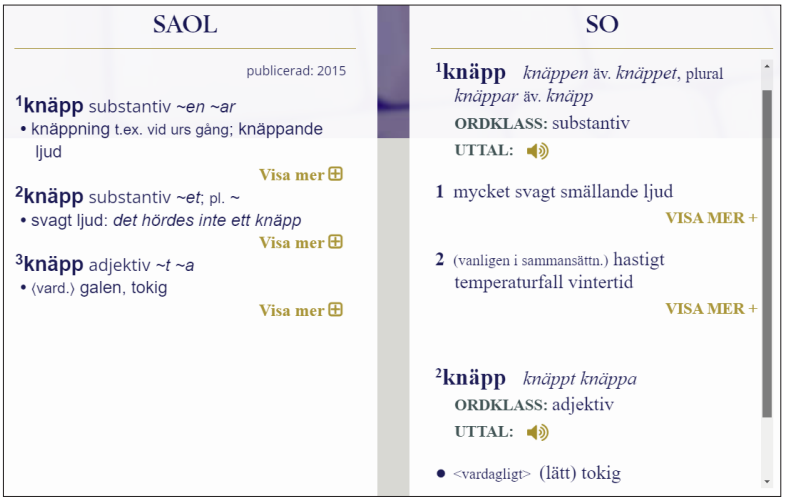
1.2. Skillnader i lemmauppdelning mellan SO och SAOL

Från att ha varit endast tryckta böcker som användarna har kunnat slå i var för sig, finns sedan 2017 samtidsordböckerna SAOL och SO, tillsammans med den (framförallt) historiska SAOB, på webbplatsen Svenska.se, där de visas upp bredvid varandra och således enkelt kan jämföras. Detta leder till att ordboksredaktion och användare enkelt kan upptäcka skillnader, inkonsekvenser etc. i ordböckernas mikrostrukturer. Redaktionen för SAOL och SO, till vilken artikelförfattaren hör, arbetar därför nu bland annat med harmonisering av nämnda ordböcker i syfte att minimera omotiverade innehållsliga skillnader mellan dem. Observera

”omotiverade”: det finns nämligen gott om *motiverade* skillnader mellan samtidsordböckerna, t.ex. att SAOL är mer normativ i fråga om böjning (exempelvis att *bikini* bör böjas *bikinier* i plural), medan SO är mer deskriptiv (*bikini* kan böjas *bikinis* i plural); att betydelsebeskrivningarna skiljer sig åt, där definitionsordboken SO generellt ger utförligare beskrivningar; och att SAOL innehåller ordklassen ”namn”, vilket SO inte gör, vilket är en hävdvunnen, och här därmed räknad som motiverad, skillnad.

En *omotiverad* skillnad mellan SAOL och SO är däremot den synbarligen skiljaktiga tillämpningen av lemma-lexemmodellens principer i vissa fall, här främst den centrala princip som lyfts upp i SO (2009:IX): att två ord med olika böjning, och i och med detta typiskt tillhörande olika ordklasser, ska utgöra två lemman. SAOL ger t.ex. tre uppslagsord för *stopp*: substantivet [1]**stopp** (best. sing. *stoppet* plur. *stopp*), substantivet [2]**stopp** (*stoppen stoppar*) och interjektionen [3]**stopp** (ingen böjning). Utifrån lemma-lexemmodellen är detta förväntat: olika böjning och/eller ordklasser medför olika uppslagsord, lemman. SO har dock i detta fall en skiljaktig lemmauppdelning: interjektionen, motsvarande [3]**stopp** i SAOL, och det ena substantivet, [2]**stopp** i SAOL, saknas i SO. I stället har substantivet [1]**stopp** i SO (med böjningen *stoppet stopp*) försetts med en underbetydelse inledd med kommentaren ”äv. med funktion av en sorts interjektion”. Denna senare lösning innebär principiellt att ordet i yttrandet *Stopp!* får betraktas som ett substantiv – trots att ordet till skillnad från substantivet *stopp*, som utgör [1]**stopp** i SO, är oböjligt.

Ett annat exempel, där SAOL återigen förefaller följa lemma-lexemmodellen striktare än SO, gäller **knäpp**. Låt oss bortse från adjektivlemmat och i stället fokusera på substantivlemmana. I SAOL delas som synes [1]**knäpp** och [2]**knäpp** upp, vilket torde kunna motiveras av deras olika böjningar: [1]**knäpp** böjs i best. sing. *knäppen* och i plur. *knäppar*, medan [2]**knäpp** böjs *knäppet knäpp*. SO väljer dock att sammanföra **knäpp** *knäppen knäppar* och **knäpp** *knäppet knäpp* i ett lemma, [1]**knäpp**. Se figur 2.



FIGUR 2. Uppslagsordet **knäpp** i SAOL och SO på Svenska.se.

Som synes innehåller SAOL och SO delvis skilda betydelser av substantivet *knäpp*. Det ljud som beskrivs i **¹knäpp** ”t.ex. vid urs gång” i SAOL har en motsvarighet i SO som beskrivs som ett ”smällande” ljud (exemplet i SO är *det hördes en knäpp när låset gick igen*), och **¹knäpp 2** i SO ’hastigt temperaturfall vintertid’ verkar inte vara representerat i den aktuella artikeln i SAOL. Centralt här är i alla händelser att SO sammanför två böjningsmönster i ett lemma. Detta är i sin ordning om det råder ”fri” variation (jfr exemplet med *svanar* och *svanor* i 1.1 ovan), men frågan är om den är så fri här. I betydelsen ’mycket svagt smällande ljud’, **¹knäpp 1** i SO, anges exempel med utrum, *en knäpp*, i SO. Betydelsen kan av allt att döma även böjas i best. sing. som *knäppet*, vilket motsvaras av böjningsinformationen till SAOL:s **²knäpp**. Således är det fri variation så långt.[4] Problemet uppstår vid **¹knäpp 2** i SO, ’hastigt temperaturfall vintertid’, där det närmast verkar vara sammansättningen *köldknäpp* som avses. Sammansättningen **köldknäpp** böjs emellertid enligt SO enbart *köldknäppen köldknäppar* (dvs. inte *köldknäppet köldknäpp*) och det finns ingen angivelse av att något annat skulle gälla vid användning i simplexformen *knäpp*. Detta är en indikation om att lexemet **¹knäpp 2** i SO nog borde göras till lemma om lemma-lexemmodellen ska avspeglas i artikeln.

4 Som noteras av en granskare, är dock idiomen till **¹knäpp 1**, t.ex. **en knäpp på näsan**, knappast möjliga att byta genus på: jfr *?ett knäpp på näsan*.

Det ska sägas att vissa lexikaliska principer faller inom ramen för lemma-lexemmodellen, beroende på vilka ordklasser man antar. En sådan är att inskränka en betydelse till en viss böjningsform i ordets böjningsmönster, t.ex. i en viss syntaktisk funktion. Adjektivets *t*-form kan nämnas som exempel. *Svenska Akademiens grammatik*, SAG (2:626), anger att *t*-formen av adjektiv betraktas som en böjningsform av adjektivet snarare än som en avledning till adverb: ”I denna grammatik betraktas formerna på *-t* inte som adverb utan som adjektiv också när de står adverbiellt”. Detta verkar vara i linje med SO:s generella hantering av *t*-former, en spegling av principen att inte innehålla ”*t*-avledda adverb som egna lemman och lexem” som tillämpades redan i uppbyggnaden av den databas som SO härstammar från (Järborg 1996:39). SAOL följer emellertid ett, som det verkar, mer traditionellt ordklassbruk och ger t.ex. lemmastatus för såväl adjektivet **hemsk** som för *t*-formen, här adverbet, **hemskt**. SO placerar sig alltså närmare SAG i att ge adjektivet **hemsk** med *t*-formen i underbetydelsen ”i adverbiell användning äv. som rent förstärkningsord”, t.ex. *det var hemskt roligt att träffas*. Det bör diskuteras hur konsekvent ordklasser delas in i de två ordböckerna i förhållande till SAG, men det är i de här redovisade fallen åtminstone inte fråga om några påtagliga avvikelser från lemma-lexemmodellens grundläggande principer: en viss stavning med en viss ordklass och ett visst böjningsmönster redovisas som ett lemma.

2. Dubbla ordklasser i SAOL

SAOL:s princip är enligt inledningen (SAOL 14:XVIII) att ”Likstavade ord med olika böjning betraktas som olika ord och ges med framförställd indexeringssiffra”, vilket kan sägas beskriva implementeringen av (åtm. en del av) lemma-lexemmodellen. Samtidigt noteras i samma inledning (s. XVII) att ett antal fall av dubbla ordklasser ändå förekommer i SAOL. Detta medför avsteg från nämnda modell: olika ordklasser borde utgöra olika lemman, vilket de gör i SO.

De aktuella dubbla ordklasserna innefattar ofta ordklasser som böjs på samma sätt; typiskt är de oböjliga. Lemmat **inifrån**, med dubbelordklassen ”preposition och adverb” i SAOL (i SO bildar preposition och adverb skilda lemman), är t.ex. oböjligt oavsett om det används som preposition (t.ex. i frasen *inifrån huset*) eller som adverb (t.ex. i *dörren öppnas inifrån*). Medlemmarna i den relativt stora dubbelordklassen ”adverb och

adjektiv" följer dock inte samma mönster. I denna grupp återfinns många ord med *-vis*, t.ex. **delvis** och **fläckvis**, som i adverbanvändning är oböjliga men i adjektivanvändning kan kongruensböjas *-vist* och *-visa*. Här rymmer samma lemma alltså såväl olika ordklasser som olika böjning, vilket kan ses som avsteg från lemma-lexemmodellen som kräver att användaren på förhand vet att adverblemmat är oböjligt (i artikelhuvudet ges endast *adjektivets* böjning med *-vist -visa*). I den utökade böjningsinformationen i den elektroniska SAOL anges förvisso endast en böjningsform, uppslagsformen, för adverbet, vilket underlättar om användaren förstår hur den ska tolka böjningsparadigmet. I SO är adverb och adjektiv åtskilda lemman, så där uppstår inte just nämnda problem. Men det kan såklart se anmärkningsvärt ut för användaren med olika lemmalösningar i de olika verken.

3. Ordklassövergångar

Ordklassbestämning i ordböcker har rönt ett visst intresse och har inte sällan utgått från enskilda intressanta fall. Bland studier i en nordisk kontext kan Martola (2013) och Jensen (2013) nämnas. Få studier verkar emellertid ha undersökt ordklassbestämning ur ett harmoniseringsperspektiv och utifrån en teoretisk lexikologisk modell såsom lemma-lexemmodellen.

I detta avsnitt ges några exempel på ordklassövergångar, som avser fall där en ordbok kan sägas frångå lemma-lexemmodellen genom att i en ordboksartikel om ett lemma av en viss ordklass infoga inskränkningar av en betydelse som hänför sig till en annan ordklass än lemmats. SAOL innehåller, såvitt jag kan se, inte några tydliga fall av dessa övergångar. I SO kan man däremot finna flera exempel på sådana, och jag har i redaktörsgränssnittet fritextsökt efter ordklasser i de informationsfält som innehåller formkommentar och inledare till underbetydelse för att finna relevanta fall.

Exempel på övergångar mot particip, uteslutande från verb (se vidare 3.1), erhålls genom (wildcard-)sökning efter "part[icip]". Den andra ordklassen jag har sökt efter är substantiv (se 3.2), genom sökning efter "substantivisk" och "substantiverat". Övergångarna är till klart övervägande del från adjektiv (237 av 265 fall), men det går också att identifiera en grupp av övergångar (se 3.3) från interjektioner (12 av 265 fall). Övriga fall är av blandad karaktär och undersöks inte vidare här. Sökning efter

”interjektion” ger få fall av övergångar, uteslutande från substantiv mot ”funktion som en sorts interjektion” och liknande kommentarer (se 3.3).

Det förekommer i övrigt främst ett fåtal fall av prepositionell användning och användning som subjunktion (i båda fallen av adverb). Dessa lämnas av utrymmesskäl utanför undersökningen.

3.1. Verb och participiella användningar

I SAG anges i participkapitlet att ”Participen är *avledningar* av verb” (SAG 2:582, min kursivering). Med denna beskrivning anges att particip i SAG utgör en ordklass och inte, som traditionellt ofta har antagits, böjningsformer av verb (t.ex. Jörgensen & Svensson 1987:29). Participen är å ena sidan verbala bl.a. i att de inte sällan har liknande valens som de verb de är avledda ifrån. Å andra sidan fungerar participen ofta syntaktiskt likt adjektiv, t.ex. i *Cykeln är stulen*.

I SAOL och SO saknas particip som ordklass på lemmanivå; i stället behandlas de främst som böjningsformer av lemmats ordklass (verb) och som adjektiv. I SAOL är t.ex. *avliden* endast en perfekt participform till verbet **avlida**. I SO ges emellertid en särskild artikel för **avliden** som adjektiv. Det omvända gäller t.ex. för **renrakad**, ett adjektiv i SAOL, men i SO endast ett verb, **renraka**, med kommentaren ”nästan enbart perfekt particip” (**renraka 1**) resp. ”vanligen perfekt particip” (**renraka 2**). Detta är oväntat, med tanke på att artikeln **particip** i SO beskriver particip som en ’*avledning* av verb som kan ha samma funktioner som ett adjektiv’ (min kursivering). Om particip är en egen ordklass, borde det i princip ha lemmastatus.

Presens particip hanteras på liknande sätt: i SAOL och SO saknas ett uppslagsord **nydanande**; i stället behandlas ordet som en participiell böjningsform under verbet **nydana** i SAOL, i SO tillsammans med formkommentaren ”vanligen presens particip”. På liknande sätt befinner sig *nekande* i t.ex. *en nekande sats* under verbet **neka 2** i SO, tillsammans med formkommentaren ”nästan enbart presens particip”. Motsvarande participform är ett (oböjligt) adjektivlemma i SAOL.

Inledningen i SO 2009 (s. XII) anger att ”Uppsättningen ordklasser sammanfaller med den traditionella […]”, och även om ordklassuppsättningen inte redovisas i SAOL, finns det mycket som tyder på att SAOL i stort sett följer SO:s ordklassuppsättning (med vissa undantag,

t.ex. SAOL:s ordklass ”namn”). I de aktuella fallen gäller saken alltså främst vilken status *particip* har och bör ha: om de ska fortsätta räknas som böjningsformer av verb faller nuvarande behandling av dem i stort sett in i lemma-lexemmodellen. Om SO ämnar följa sin beskrivning av **particip** som en *avledning* av verb, och om man har som ambition att följa SAG:s ordklassindelning, borde dock particip nog utgöra egna lemman. Sedan uppstår frågan om denna hantering borde harmoniseras med SAOL:s.

3.2. Adjektiv, particip och substantiviska användningar

Liksom gränsen mellan particip och verb är gränsen mellan adjektiv, particip och substantiv dragen på olika sätt i SAOL, SO och SAG.

Adjektiv kan enligt SAG användas i nominalfraser med adjektiviskt ”huvudled” (SAG 3:249), som i *Jesus botade halta och lytta*. I linje med detta är **lytt** i SAOL och SO endast klassat som adjektiv. I SO ges även tillägget ”ofta substantiverat”, vilket kan tolkas på flera sätt. SAG (1:228) anger att *substantiverad* kan avse minst två saker: ”ibland om adjektiv som genom *ordklassbyte övergått till substantiv (t.ex. *liberal-en*) eller som används *självständigt (t.ex. *mina bekant-a*)”. Om det handlar om det första, att adjektivet genomgått ordklassbyte, frångås i princip lemma-lexemmodellen, men eftersom böjningen följer adjektivets (*lytt*, *lytta* osv., inte t.ex. **lytten*) kan det antas att det handlar om självständig användning i nominalfras, vilket väl får sägas följa lemma-lexemmodellen.

I fråga om adjektiv avledda av particip är situationen något mer komplicerad: ett ord som *åtalad* i *många åtalade i muthärvan* återfinns som participiell böjningsform under verbet **åtala** i SAOL. I SO slås det emellertid upp som ett adjektivlemma, **åtalad**, med böjningsangivelsen *åtalat åtalade* och med formkommentaren ”ofta substantiverat”. Ett problem för ordboksanvändaren blir här att förstå att den angivna neutrumformen *åtalat* näppeligen används i substantiverad variant (jfr ??*mycket åtalat i muthärvan*).

3.3. Interjektioner och substantiviska användningar

SAG redovisar interjektionsliknande ord, t.ex. *ja* och *nej*, i interjektionskapitlet som just interjektioner, samtidigt som det anges att användning

av interjektionens uttryck gör att interjektionen tenderar att ”lexikaliseras som substantiv” (SAG 2:768). I såväl SAOL som SO separeras ofta ord som *ja* och *nej* i olika lemman för interjektion och substantiv, där interjektionen är oböjlig medan substantivet är böjligt (t.ex. *jaet* i bestämd form singular). Hanteringen av exklamativa uttryck som *Stopp!*, som berördes tidigare, är dock annorlunda i ordböckerna: SAOL ger [1]**stopp** (subst., *stoppet*, *stopp*), [2]**stopp** (subst., *stoppen*, *stoppar*) och [3]**stopp** (interj.), medan SO bara ger [1]**stopp 1** (subst., *stoppet*, *stopp*) med underbetydelsen ”äv. med funktion av en sorts interjektion”: *”Stopp där!” sade vakten*, vid sidan av [2]**stopp** (adv.). Detta innebär att *Stopp!* i SAOL klassas som interjektion, liksom i SAG, och som substantiv i SO, ”med funktion av en sorts interjektion”, samtidigt som lemmats böjning anges vara ”*stoppet*, plural *stopp*, bestämd plural *stoppen*”.

Ett slags omvänt förhållande mellan SAOL och SO förekommer också. I SAOL är t.ex. **mums** en interjektion ([1]**mums**), men även ett substantiv ([2]**mums**) som illustreras med frasen *smaka mums*. I SO är **mums** i *det smakade mums* emellertid endast interjektion, med exemplet *Pannkakor! Mums!*, och underbetydelsen ”äv. i adverbiell och substantivisk användning”, med exempel som *det var mums det!*. Detta innebär att *mums* i *smaka mums* i SAOL klassas som substantiv, och därmed med böjningsformer som den bestämda formen *mumset*, medan *mums* i SO endast klassas som interjektion (”i adverbiell och substantivisk användning”). Eftersom lemmat har interjektionsböjning, saknar SO substantivets bestämda form *mumset* som ges i SAOL. (I SAG är *mums* förvisso endast behandlat i interjektionskapitlet, men *mums* i *smaka mums* torde syntaktiskt vara svårt att avgränsa från adverb.)

I samtliga fall gäller att interjektionsanvändningarna i substantivartiklarna på ett sätt avviker från lemma-lexemmodellen, vilket får som konsekvens att böjningsangivelsen (t.ex. för **stopp** med best. form *stoppet*) i princip felaktigt kommer att gälla även för interjektionen i uttrycket *Stopp!*.

4. Mot en harmonisk lemma-lexemmodell – vilka ordklasser ska vi ha?

Sammanfattningsvis visar genomgången att lemma-lexemmodellen i flera fall följs i mycket olika utsträckning och på olika sätt i SO och SAOL

och att ordklassövergångarna förtjänar särskild uppmärksamhet i kommande studier. I vissa fall, t.ex. i fråga om substantiveringar av adjektiv och substantiveringar av interjektioner, kan det i alla händelser finnas anledning att bryta ut ordklassövergångarna till egna lemman, vilket skulle kunna göra dem lättare att finna av användarna. Dessutom skulle böjningsmönstren tydliggöras, och en utbrytning skulle i många fall bidra till harmoniseringen av ordböckerna.

En viktig fråga som aktualiseras av genomgången av lemma-lexem-modellen i harmoniseringsarbetet är som sagt vilken ordklassuppsättning som ska antas i ordböckerna. Om SAG:s uppsättning (och kriterier för denna uppsättning) skulle väljas skulle det krävas mycket stora insatser. Exempelvis skulle då hela participklassen behöva omarbetas, på ett visst sätt i SO och på ett delvis annat sätt i SAOL. Det är dock svårt att undvika det faktum att tre språkliga referensverk som behandlar samtida svenska, SAG, SO och SAOL,[5] har samma avsändare, så det är kanske inte orimligt att som användare av dem begära att de är någorlunda samstämmiga och, om så inte är fallet, att skillnaderna kan motiveras. Det är som synes en viktig fråga, som inte ofta behandlats i den lexikografiska litteraturen.[6]

Referenser

Allén, Sture 1999 [1981]. The Lemma-Lexeme Model of the Swedish Lexical Database. *Modersmålet i Fäderneslandet. Ett urval uppsatser under fyrtio år av Sture Allén*. (Meijerbergs arkiv för svensk ordforskning 25). Göteborg: Meijerbergs institut för svensk etymologisk forskning, 268–278. (Publicerad första gången 1981.)

Borin, Lars 2008. Lemma, lexem eller mittemellan? Ontologisk ångest i den digitala domänen. I: Jóhannesson, Kristinn et al. (red.), *Nog*

5 Eftersom SAOB har ett presentationssätt grundat i andra principer än SAOL och SO och är en ”historisk och samtidsspråklig ordbok” snarare än ett renodlat samtidsverk (se Nilsson & Rosqvist 2022:53, 68), väljer jag att inte inkludera SAOB här, även om visst harmoniseringsarbete såklart torde kunna inkludera även denna ordbok.

6 Se t.ex. Svensén (2004:179–185), som i kapitlet om ordklasstillhörighet grundligt går igenom bl.a. syftet med ordklasser i en ordbok men inte berör frågan om på vilka grunder eller utifrån vilka grammatiska beskrivningar ord i en ordbok ska tilldelas en viss ordklass.

ordat? Festskrift till Sven-Göran Malmgren. Göteborg: Meijerbergs institut för svensk etymologisk forskning, 59–67.
Jensen, Eva Skafte 2013. Ordklasseproblemer, tilfældet *sådan*. *LexicoNordica* 20, 55–74.
Järborg, Jerker 1989. Betydelseanalys och betydelsebeskrivning i Lexikalisk databas. (Preliminär version). Göteborg: Göteborgs universitet.
Järborg, Jerker 1996. Formaliserad lexikologi. Rapport från ett långtidsprojekt. (Preliminär version). Göteborg: Göteborgs universitet.
Jörgensen, Nils & Jan Svensson 1987. *Nusvensk grammatik*. Malmö: Liber.
Malmgren, Sven-Göran 2014. Svenska Akademiens ordlista genom 140 år: mot fjortonde upplagan. *LexicoNordica* 21, 81–98.
Martola, Nina 2013. Mediala s-verb i svenska ordböcker. *LexicoNordica* 20, 93–109.
Nilsson, Pär & Bodil Rosqvist 2022. Nya tider, nya möjligheter – inför en reviderad version av SAOB i en helt ny tid. *LexicoNordica* 29, 53–72.
Ralph, Bo, Jerker Järborg & Sture Allén 1977. Svensk ordbok och Lexikalisk databas. Förstudierapport. Göteborg: Göteborgs universitet.
SAG 1–4 = Teleman, Ulf, Staffan Hellberg & Erik Andersson 1999. Del 1–4. *Svenska Akademiens grammatik*. Stockholm: Svenska Akademien.
SAOB = *Ordbok över svenska språket, utgiven av Svenska Akademien*. 1898–. I: <svenska.se>. Hämtat september 2022.
SAOL = *Svenska Akademiens ordlista över svenska språket*. I: <svenska.se>. Hämtat september 2022.
SAOL 14 = *Svenska Akademiens ordlista över svenska språket*. (Upplaga 14, 2015). Stockholm: Svenska Akademien.
SO = *Svensk ordbok utgiven av Svenska Akademien*. I: <svenska.se>. Hämtat september 2022.
SO 2009 = *Svensk ordbok utgiven av Svenska Akademien*. (Upplaga 1, 2009). Stockholm: Svenska Akademien.
Svensén, Bo 2004. *Handbok i lexikografi. Ordböcker och ordboksarbete i teori och praktik*. 2 uppl. Stockholm: Norstedts.

Mineralitet som leksikografisk utfordring: maskinlæring som tilnærming til semantikken

Koenraad De Smedt & Ole Martin Skilleås

It is often a challenge to define the semantics of domain specific words in dictionaries. *Mineralitet* ('minerality') and other word forms containing *mineral*, which are increasingly used as wine descriptors, are not properly defined in the most used Norwegian dictionaries. We tackle this challenge by applying machine learning algorithms to a corpus of texts describing the smell and taste of wines. A computational analysis by machine learning of context words in two classes of texts, those containing forms of *mineral* and those without, reveal some semantic fields which discriminate between the classes. We argue that such information from specialized data sources may contribute to dictionary definitions when other approaches are less feasible.

KEYWORDS: minerality, wine, Norwegian, machine learning, lexicography, meaning

1. Innledning og problemstilling

Det er ofte en utfordring å forklare fagspesifikke ord i ordbøker. Spesielt ord som er kjent blant en bred brukergruppe, men som ikke er lett å tyde presist, kan gi leksikografen hodebry. Det Norske Akademis ordbok (NAOB) har 75 ordbetydninger som har merkelappen *ønologi* (vindyrking, vinfremstilling og vinsmaking). Mens noen av disse er av relativt teknisk art, f.eks. *botrytisere, kupasje* og *edelråte*, har andre antagelig et bredere nedslagsfelt, f.eks. lånordet *brut*, den nye sammensetningen *økovin*, og neonymet *mineralitet*. Særlig for de ofte brukte ordene er det en utfordring å definere fagspesifikke betydninger på en måte som er korrekt og samtidig lett forståelig for ordboksbrukerne.

Mineralitet og *mineralsk* er blitt vanlige norske ord for å beskrive lukt, smak og andre egenskaper av noen viner. Bruken er illustrert i eksempel (1), en anmeldelse fra vintidsskriftet Vinforum.

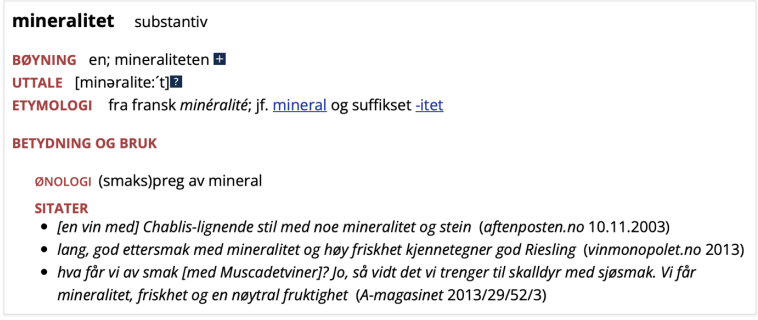
(1) [...] hvitviner med høy naturlig syre og relativt kjølige fruktaromaer samt en underliggende mineralitet fra det vulkanske jords-

> monnet. [...] Frisk, ren, ueiket og mineralsk etnavin med preg av epler, sitrus, urter og blomster. (Ronold & Johnsen 2022)

Betegnelsene *mineralitet* og *mineralsk* var knapt i bruk, hverken på norsk eller andre språk, så sent som i 2002, da Ann C. Noble tok copyright på sitt Aroma Wheel for vin, der *mineral* ikke var å finne (Noble mfl. 1984, 1987). I løpet av de siste tiårene har bruken derimot økt såpass mye at den vanlige vinmonopolkunden kan se seg tjent med en forklaring av disse ordene i vanlige oppslagsverk.

I nettversjonen av Bokmålsordboka og Nynorskordboka finnes ikke *mineralitet*, mens *mineralsk* har tautologiske definisjoner, for eksempel "som gjelder mineralene" (jf. figur 1). Disse kan ikke sies å bidra til en bedre forståelse av vinaroma. En noe mer målrettet, men lite korrekt definisjon, "(smaks)preg av mineral", finner vi for *mineralitet* i NAOB, ledsaget av autentiske eksempler som pragmatisk bidrar til betydningsavklaring (jf. figur 2).

FIGUR 1. *mineralsk* i Bokmålsordboka og Nynorskordboka (Ordbøkene.no).



FIGUR 2. *mineralitet* i Det Norske Akademis ordbok (NAOB).

Oppfatningen om at druer i mineralrik jord gir viner der man kan smake mineralene, er en myte (Maltman 2013). Mineraler er anorganiske stoffer som forekommer i ørsmå mengder i vin og de er ikke volatile – de kan ikke bli direkte tilgjengelige for olfaksjon. Likevel gir *mineralsk* som overført begrep mening blant vinsmakere. Jordsmonn er utvilsomt viktig for vinstokkenes utvikling. Vinprodusenter har en økende tendens til å navngi viner etter spesielle jordsmonn, slik som Basalt, Kalkstein, Buntsandstein osv. Geologi har dermed blitt en metafor for en kompleks smaksprofil som man forventer fra gode viner med en viss karakter eller fra visse områder, som følgende sitater antyder.

> 'Mineral' may be both descriptive of a feature of the wine, and a term for praise for wines that are expected to display this characteristic – such as Chablis. (Burnham & Skilleås 2012:20)

> Some recent research indicates that 'mineral' is a higher-order feature of the wine, and not at the same level as more straightforward elements such as fruit flavours. Thus 'mineral' would designate a set of discrete sensory elements; it would be a *character* of the wine, rather than a single element. (Burnham & Skilleås 2012:76)

> Les vins de Chablis sont des vins blancs secs qui se distinguent par leur pureté, leur fraicheur, leur finesse, leur minéralité. (Leroyer & Høy 2016:292)

Tidligere undersøkelser blant vinprodusenter og konsumenter har antydet at mineralitet til dels er knyttet til område, jordsmonn og druesort, og til dels til en smaksprofil som inkluderer bl.a. friskhet, tørrhet (syrlighet), sjø og andre aromaer. Vi kommuniserer ikke bare for å uttrykke våre preferanser, men til dels for å lede og spisse andres estetiske persepsjoner gjennom en slags sosial triangulering (Skilleås & Burnham 2012; Teil 2019). På denne måten har vinkjennere kommet til en viss konsensus om hvorvidt en vin er mineralsk, men det er fortsatt uenighet om hvilke sensoriske egenskaper som forbindes med mineralitet som sådan. Det er en viss enighet om hvilke *andre* egenskaper som forbindes med *mineralsk* (Parr mfl. 2018; Rodrigues mfl. 2015). Dette gir håp om en kontekstbasert tilnærming for å lage en brukbar definisjon av *mineralsk* og *mineralitet*.

> Results of reviewed studies overall demonstrate marked variability in both wine professionals and wine consumers' definitions and sensory-based judgments of minerality in wine, although there is some consensus in terms of the other wine attributes that associate with the term mineral (Parr mfl. 2018).

2. Data og metode

Forsøk på en mer presis definisjon av *mineralsk* og *mineralitet* er dermed en interessant leksikografisk utfordring. Med utgangspunkt i Firths (1957:11) utsagn *You shall know a word by the company it keeps*, har vi prøvd å komme litt nærmere gjennom en analyse av andre ord som opptrer sammen med *mineral, mineralsk* og lignende i beskrivelser av vinens lukt og smak. Ved hjelp av en metode basert på maskinlæring har vi testet hvorvidt man på en objektiv måte kan lære å kjenne igjen beskrivelser av mineralske viner bare ut fra kontekstord, og hvilke kontekstordene som er mest relevante i denne prosessen.

Et søk i Leksikografisk bokmålskorpus (Fjeld, Nøklestad & Hagen 2020) etter *mineralsk* gir kun irrelevante treff, f.eks. *mineralsk fosfat, mineralsk olje* og *mineralsk isolasjonsmateriale*, mens *mineralitet* gir ikke noen treff i hele tatt. Aviskorpuset (Andersen & Hofland 2012) har relevante forekomster men også irrelevante. Oppbygging av et spesialisert korpus er inntil videre uten rekkevidde, men en annen spesialisert og homogen kilde med eksempler er tilgengelig. Denne kilden er Vinmonopolets omfattende liste over alle vinene med produkttype, beskrivelser av lukt- og smak og annen informasjon (Vinmonopolet 2020). Dessverre sluttet Vinmonopolet i mellomtiden å gjøre sin liste tilgjengelig for allmennheten, så vi har brukt den siste versjonen som var tilgjengelig for oss.

Vinmonopolets liste har 21 717 produkter i 59 ulike varetyper. Deriblant er det 16 081 produkter i følgende varetyper som vi vurderer som interessante: hvitvin, champagne, annen musserende vin, perlende vin og rødvin. For hvert produkt er det 44 kolonner med informasjon men vi har brukt bare *lukt* og *smak* i analysen. Det er altså disse tekstene som utgjør vårt korpus. Tekstene er pragmatisk sett ganske uniforme: de er ikke evalueringer, men skal gi konsumenten holdepunkter for å danne seg et bilde av vinens smaksprofil.

Relevante trekk i vår analyse er nøkkelord som gjengir mest mulig konkrete sensoriske beskrivelser. Tekstene i kolonnene *lukt* og *smak*, som ofte overlapper, ble slått sammen. En viss normalisering var nødvendig for å redusere tilfeldigheter og oppnå generalisering. Eksempelvis ble *bittert* og *bitterhet* redusert til *bitter*. Alle bokstaver i tekstene ble også redusert til små bokstaver og skrivefeil i kontekstord (men ikke i *mineralsk* o.l.) ble rettet i den grad de ble oppdaget.

Tekster som inneholdt ord med *mineral*, jf. tabell 1, ble betraktet som 'mineralske tekster'. Disse ordene ble fjernet fra tekstene, samtidig som disse tekstene ble husket som tilhørende den 'mineralske' klassen, i motsetning til tekstene som ikke inneholder slike ord og derfor ble tilordnet den 'ikke-mineralske' klassen. Målsetningen var altså å se om det er mulig å finne forskjellen mellom klassene gjennom en blind prosess, dvs. når man ser bort fra ord med *mineral*.

TABELL 1. Ord som inneholder *mineral*.

kalkmineraler	kalkmineralitet	kalkmineralsk
kalkmineralske	mineral	mineralaromaer
mineralbitt	mineralbitter	mineraldominert
mineraldreven	mineraldrevet	mineralduft
minerale	mineralene	mineraler
mineralfrisk	mineralfriskhet	mineralisk
mineralitet	mineralkarakter	mineralkonsentrasjon
mineralpreg	mineralpreget	mineralrik
mineralrike	minerals	mineralsk
mineralske	mineralt	mineraltone
mineraltoner	mmineraler	rødskifermineraler
røykmineraler	saltmineral	saltmineraler
saltmineralitet	saltmineralsk	sitrusmineralitet
sjømineral	sjømineraler	sjømineralitet
sjømineralsk	skifermineraler	steinmineral
steinmineraler	steinmineralitet	steinmineralsk
vulkanmineraler		

Det annoterte korpuset ble matet inn i en samling algoritmer for maskinlæring (Scikit Learn). Tekstene ble vilkårlig fordelt i et treningssett og

et testsett. Ord som er irrelevante for vårt formål ble fjernet; disse er stoppord som *av, mot, så, annet,* osv. men også generelle ord som ikke uttrykker smak eller lukt, for eksempel, *aroma, bouquet, preg, innslag, markert, ørlite*, osv. Denne preprosesseringen gjør at en tekst som (2) blir transformert til en rekke nøkkelord i (3).

(2) sval og fruktig aroma preget av grønt eple, nesle og stikkelsbær. ung, saftig og slank, preg av sitrus, grønt eple og nesle, hint av ~~mineraler~~ i ettersmaken.

(3) sval fruktig grønt eple nesle stikkelsbær ung saftig slank sitrus grønt eple nesle

Med disse dataene ble det konstruert ulike semantiske modeller. Algoritmene finner selv ut i hvilken grad de ulike ordene i treningssettet er relevante for forskjellen mellom de to tekstklassene. Selvsagt er det interessant for oss å få vite hvilke ord algoritmene vurderte som de mest karakteristiske for den ene eller den andre klassen.

3. Resultater

De ulike algoritmene har hver sine egne parametere; noen legger litt mer vekt på frekvente ord, mens andre legger litt mer vekt på mindre frekvente ord. Likevel et det betraktelig overlapp i resultatene. Vi vil derfor beskrive én modell laget av én algoritme, Stochastic Gradient Descent med L1 og L2 regularisering (Scikit Learn SDG).

Ved testing gav denne modellen 78 % korrekte resultater, som er sammenlignbart med de andre modellenes ytelse. Dette viser at kontekstord bidrar med informasjon, men denne prestasjonen ikke skal overdrives, siden klassene er skjevt fordelt: 72 % av tekstene har ikke ord med *mineral*. Mer interessant i vår sammenheng er kontekstordene med høyest diskriminerende verdi for den ene eller den andre klassen, som vist i tabell 2.

TABELL 2. Noen karakteristiske ord i smaksnotater uten og med *mineral*. Jo høyere verdien er, desto mer karakteristisk er ordet for 'mineralske' viner.

-3.665	*umiddelbar*	3.191	*sitrus*
-2.083	*fedme*	2.235	*sjøgress*
-1.991	*bitter*	2.234	*boysenbær*
-1.906	*knust*	2.050	*kritt*
-1.814	*drops*	1.778	*lime*
-1.751	*søt*	1.731	*kalkholdig*
-1.605	*druer*	1.688	*edel*
-1.563	*sukker*	1.661	*transparent*
-1.547	*perlende*	1.655	*mandarin*
-1.519	*nøtt*	1.610	*flint*
-1.506	*appelsin*	1.605	*mynt*
-1.365	*karamell*	1.531	*stein*
-1.361	*fruktsødme*	1.515	*slank*
-1.343	*pære*	1.488	*rem*
-1.245	*undermoden*	1.467	*vekt*
-1.222	*jordmonn*	1.404	*gul*
-1.214	*tørket*	1.289	*kjølig*
-1.198	*trekrydder*	1.266	*eplekart*
-1.170	*behagelig*	1.225	*salt*
-1.162	*toast*	1.209	*sjø*
-1.111	*tjære*	1.175	*fullmoden*
-1.085	*tannin*	1.126	*autolyse*
-1.085	*balsam*	1.124	*sjøsalt*
-1.081	*banan*	1.119	*kjeks*
-1.071	*ingefær*	1.119	*kompleks*

Tre grupper av semantisk relaterte sensoriske kontekstord skiller seg ut: *sitrus/lime/mandarin*, *sjø/sjøgress/(sjø)salt*, og *kritt/kalk/flint/stein*. Disse resultatene kan gi holdepunkter til en mer konkret, sensorisk definisjon av *mineralsk*. Kontekstord som *edel*, *transparent*, *slank* og *kompleks* er positive og antyder at aromaene er fine og subtile; dermed er mineralitet trolig medvirkende til en oppfatning av kvalitet og raffinert eleganse. Det kan også være fristende å definere ordet i negative termer: *mineralsk* kan betraktes som et generelt stiltrekk i vin, som omtrent diametralt mot-

satt av *fruktig* og *umiddelbar*. Ikke-mineralske viner har ofte en uttalt smak (*bitter, søt*) og mindre subtil lukt (*drops, nøtt, karamell, pære, tjære, banan, ingefær*).

4. Diskusjon

Resultatene peker i en retning som er mer lovende enn definisjonene i dagens norske ordbøker. Før vi kommer med et forslag i denne retningen, må vi likevel ta noen forbehold. Tilfeldigheter kan gjøre at noen ord forekommer hyppigere i den ene kategorien enn i den andre, uten at det er av betydning. Derfor har vi valgt en algoritme med regularisering, som til en viss grad kan glatte over tilfeldigheter. Fortsatt er det viktig å se på resultatene med et kritisk blikk og å se på de store linjene heller enn på detaljene.

I datagrunnlaget er det samlet lukt- og smaksnotater for ulike produkttyper. Dessuten er de 'mineralske' notatene ikke likt fordelt over de ulike kategoriene. Av rødvinene er ca. 11 % beskrevet som mineralske, mot 17 % av perlende, 25 % av musserende (bortsett fra champagne), 50 % av champagne og 51 % av hvitvinene. Aromabeskrivelser som *autolyse* og *kjeks* er typiske for champagne og andre musserende viner, og kan derfor være mindre relevante for ikke-musserende viner. At hele 2871 produkter, godt over halvparten av alle 'mineralske' produkter i de inkluderte kategoriene, er hvitviner, er mindre overraskende enn at en ikke-triviell andel rødviner skal være mineralske. Det kan hende at *mineralitet* tilsvarer en annen aromaprofil i de ulike kategoriene, slik at en mineralsk rødvin har en annen kompleks profil enn en mineralsk hvitvin eller champagne, men det kan også bety at ordets betegnelse har forflyttet seg ganske langt fra den typiske Chablis-profilen.

Den enorme andelen 'mineralske' viner kan tyde på en viss bleking av begrepet, en tendens som utvilsomt har blitt påvirket av vinprodusentene og -innkjøperne, som har redusert *mineralitet* til et instrument for merkevarebygging av visse viner (Leroyer 2022; Temmerman 2017). Denne kommersialiseringen bidrar sannsynligvis til at amatører som ikke har lært seg referansepunktene kan komme i skade for å bruke betegnelser man har lest seg til, uten å kunne anvende disse på en pålitelig måte. Det er derfor en viss fare for at vinsmakingsvokabularet blir fanget i en boble der bekreftelsestendenser gir overforbruk.

5. Konklusjon

I denne lille studien har vi tatt utgangspunkt i en utfordring knyttet til orddefinisjon. Riktignok er det ganske mange ord, blant annet abstrakte, som skaper leksikografiske utfordringer, og lett kan føre til sirkeldefinisjoner. *Kunst*, for eksempel, lar seg i utgangspunktet ganske korrekt, og i pakt med den utbredte institusjonelle definisjonen av kunst, definere som "alt som deltakerne i kunstinstitusjonen betrakter som kunst", uten at dette ville hjulpet ordboksbrukeren særlig mye. På samme måte kan *mineralitet* defineres som "aroma som vinsmakere kaller mineralsk", som er konsist og mer korrekt enn "(smaks)preg av mineral", men likevel ikke videre nyttig.

Noen ganger er det en vei ut ved hjelp av data, som vi har prøvd å vise i denne artikkelen. Det er etter hvert blitt vanlig for leksikografene å hente informasjon fra korpus (f.eks. Rauset mfl. 2022). I dette tilfellet vurderte vi bruk av et mer spesialisert datagrunnlag som mer effektiv enn bruk av andre eksisterende korpus. Videre har vi brukt semantisk modellering med maskinlæring basert på kontekstord, en metode som vi ikke vil fremheve som en generell metode i leksikografi, men slik modellering kan i sjeldne tilfeller supplere andre korpusbaserte metoder.

På bakgrunn av de studerte kontekstordene og det vi vet om vintypene, kan et forsøk til definisjon for *mineralitet* være følgende: *aroma og smaksprofil som ofte assosieres med sitrus, sjøgress og kritt, og som er mest karakteristisk for noen friske hvitviner.* Ved å bruke tekster fra én stor men pragmatisk ensidig kilde, som har relativt uniforme, varedeklarative men ikke-evaluerende beskrivelser av lukt og smak, har vi lagt vekt på sensoriske karakteristikker. En mulig innvending er at *mineralitet* spiller en mer mangfoldig rolle i vinanmeldelser og at begrepet står sentralt i en rekke evalueringsscenarier. Vil man gå dypere inn i hva *mineralitet* gjør med vinens samlede profil og kvaliteter i ulike typer diskurs, kan det være interessant å lage et bredere korpus av vinrelaterte tekster.

Litteratur

Andersen, Gisle & Knut Hofland 2012. Building a large corpus based on newspapers from the web. I: Andersen, Gisle (red.), *Exploring Newspaper Language: Using the web to create and investigate a large*

corpus of modern Norwegian, *Studies in Corpus Linguistics*. Amsterdam/Philadelphia: John Benjamins, 1–28.

Burnham, Douglas & Ole Martin Skilleås 2012. *The aesthetics of wine*. Chichester, UK: Wiley-Blackwell.

Firth, John Rupert 1957. *Studies in linguistic analysis*. Oxford: Blackwell.

Fjeld, Ruth Vatvedt, Anders Nøklestad & Kristin Hagen 2020. Leksikografisk bokmålskorpus (LBK) – bakgrunn og bruk. *OSLa* 11:1, 101–124. doi: https://doi.org/10.5617/osla.8176.

Leroyer, Patrick 2022. Médiatisations lexicographiques et branding du vin. I: Lavric, Eva, Cornelia Feyrer & Carmen Konzett-Firth (red.), *Le vin et ses émules: Discours œnologiques et gastronomiques*. Berlin: Frank & Timme, 535–558.

Leroyer, Patrick & Asta Høy 2016. Vinsmagningsordbogen Œnolex Bourgogne. En milepæl i pragmatisk fagleksikografi. *Nordiske studier i leksikografi* 12. Oslo: Nordisk forening for leksikografi, 287–302.

Maltman, Alex 2013. Minerality in wine: a geological perspective. *Journal of Wine Research* 24:3, 169–181. doi: 10.1080/09571264.2013.793176.

NAOB = *Det Norske Akademis ordbok*. Det Norske Akademi for Språk og Litteratur. <https://naob.no>. Hentet 22. mai 2022.

NAOB. *mineralitet*. <https://naob.no/ordbok/mineralitet>. Hentet 22. mai 2022.

Noble, Ann C., Rich A. Arnold, John Buechsenstein, E. Jane Leach, Janice O. Schmidt & Peter M. Stern 1987. Modification of a standardized system of wine aroma terminology. *American Journal of Enology and Viticulture* 38:2, 143–146.

Noble, Ann C., Rich A. Arnold, Bryce M. Masuda, Suzanne D. Pecore, Janice O. Schmidt & Peter M. Stern 1984. Progress towards a standardized system of wine aroma terminology. *American Journal of Enology and Viticulture* 35:2, 107–109.

Ordbøkene.no. *mineralsk*. <https://ordbokene.no/bm,nn/search?q=mineralsk>. Hentet 22. mai 2022.

Parr, Wendy V., Alex J. Maltman, Sally Easton & Jordi Ballester 2018. Minerality in Wine: Towards the Reality behind the Myths. *Beverages* 4:4, 77. doi: 10.3390/beverages4040077.

Rauset, Margunn, Gyri Smørdal Losnegaard, Helge Dyvik, Paul Meurer,

Rune Kyrkjebø & Koenraad De Smedt 2022. Words, Words! Resources and Tools for Lexicography at the CLARINO Bergen Centre. I: Fišer, Darja & Andreas Witt (red.), *CLARIN: The Infrastructure for Language Resources*. Berlin: De Gruyter, 457–480.

Rodrigues, Heber, Jordi Ballester, Maria Pilar Saenz-Navajas & Dominique Valentin 2015. Structural Approach of Social Representation: Application to the Concept of Wine Minerality in Experts and Consumers. *Food Quality and Preference* 46, 166–172. doi: 10.1016/j.foodqual.2015.07.019.

Ronold, Arne & Eirik Sand Johnsen 2022. Ukens vin: energi fra Etna. *Vinforum*, 16. mars. <https://www.vinforum.no/Artikler/Ukens-vin/Ukens-vin-Energi-fra-Etna>. Hentet 6. april 2022.

Scikit Learn. *Machine Learning in Python*. <https://scikit-learn.org>. Hentet 22. mai 2022.

Scikit Learn SDG. *Stochastic Gradient Descent*. <https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html>. Hentet 22. mai 2022.

Skilleås, Ole Martin & Douglas Burnham 2012. Patterns of Attention: "Project" and the Phenomenology of Aesthetic Perception. *Rivista di estetica* 51, 117–135. doi: 10.4000/estetica.1399.

Teil, Geneviève 2019. Learning to smell: on the shifting modalities of experience. *The Senses and Society* 14:3, 330–345. doi: 10.1080/17458927.2019.1665812.

Temmerman, Rita 2017. Verbalizing sensory experience for marketing success: The case of the wine descriptor minerality and the product name smoothie. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication* 23:1, 132–154. doi: 10.1075/term.23.1.06tem.

Vinmonopolet 2020. *Produkt- og butikkdata fra Vinmonopolet*. <https://www.vinmonopolet.no/datadeling>. Hentet 30. september 2020.

Eit ikon møter ein fullformgenerator. Om Ivar Aasens *Norsk Ordbog med dansk Forklaring* (1873)

Oddrun Grønvik, Christian-Emil Smith Ore & Trond Minde

Many European languages have undergone considerable orthographic changes over the last 150 years. This hampers the application of modern computer-based analysers to older text, and hence computer-based annotation and studies of text collections spanning a long period. As a step towards a functional analyser for Norwegian texts (Nynorsk standard) from the 19th century, funding was granted in 2020 for creating a full form generator for all inflected forms of headwords found in Ivar Aasen's dictionary *Norsk Ordbog med dansk Forklaring* published in 1873 and his grammar *Norsk Grammatik* from 1864.

Creating the Aasen word bank has led to new insight into the dictionary, its structure, internal organisation, and ambition level as well as its link to Aasen's 1864 grammar.

KEYWORDS: orthographic history, full form generator, text analysis, Ivar Aasen, Nynorsk

1. Innleiing

Ein fullformgenerator for tidleg nynorsk, *Norsk Ordbog med dansk Forklaring* (Aasen 1873), vart fullført sommaren 2021. Bygginga av fullformgeneratoren førte med seg eit detaljert kategoriseringsarbeid som auka innsikta i Ivar Aasens ordbokarbeid. For ein nærare omtale av fullformgeneratoren, heretter Aasen-ordbanken, sjå Ore et al. (2023).

Med Aasen-normalen forstår vi summen av ord som er leksikalske gjenstandar (Atkins & Rundell 2008:163), og er uttrykte i ortografien frå *Norsk Grammatik* (Aasen 1864) og Aasen 1873. Om lag 50 000 av dei viktigaste orda frå norsk talemål er handsama i Aasen 1873, men gjeve ordlagingssystemet for germanske språk, med samansetning, samanskriving av ordledd og eit rikt register av avleiingsaffiks, seier det seg sjølv at lemmainventaret i tekst er større.

Ein fullformgenerator krev ei lemmaliste i ønskt rettskriving, bøyingsskjema for ordtypar som krev bøying, og programvare, her same program

som i Norsk Ordbank (Ore & Grønvik 2018a). Ein truverdig fullformgenerator bygd på Aasen-normalen krev at rettskrivinga i Aasen 1873 blir registrert både for oppslagsformer, bøygde former og fleirordsuttrykk. Dette vart gjort ved å isolera ei lemmaliste frå Aasen 1873, laga paradigmeskjema ut frå Aasen 1864, og tilordna paradigme til kvart lemma. I samband med tilordninga av paradigme måtte lemma i normert form skiljast frå oppslagsformer i Aasen 1873 som ikkje representerer Aasen-normalen, og målføreformer med lemmastatus måtte få oppslagsform i samsvar med Aasen-normalen.

Men ei ordbok er ikkje ei normert lemmaliste – særskilt ikkje ei ordbok som dokumenterer ordtilfanget i eit talemålsområde for første gong. Aasen 1873 er ein svært kompakt tekst. Det einaste førehandskjende var at det ikkje før hadde vore gjort ei detaljert relasjonell kartlegging av innhaldet i ordboka. Vi rekna med at arbeidet ville gje handfast innsyn i Ivar Aasens metode og føremål. Denne artikkelen gjer greie for det vi fann om den indre organiseringa av Aasen 1873. Artikkelen følgjer prosessen, frå maskinell uthenting av lemma til analyse av artikkeltypar.

2. Uthenting av lemmaliste til fullformgeneratoren

Aasen-ordbanken skal ha med alle leksikalske gjenstandar som er handsama i Aasen 1873. Ordboksteksten i Aasen 1873 vart innskriven som elektronisk tekst i 1996, med noko forenkla typografi jamført med originalen (Kruken & Aarset 2003:XXVIII). Artikkeloverskrifter er sette i halvfeit skrift, følgde av ordklasse. Alle oppslagsord i artikkeloverskrifter vart henta maskinelt og fekk post i Ordbanken.

Inne i artikkeltekstane finst døme på avleiingar, samansetningar og fleirordsuttrykk. Forma på ordtilfanget varierer i både typografi og skrivemåte. Ein del ord i artikkelteksten er sette i halvfeit, har tilføydd ordklasse, og ser ut til å ha normert status. Desse vart også registrerte maskinelt. Den maskinelle uthentinga gav ein ordbank med drygt 42 000 postar, fordelte på om lag 36 000 ordboksartiklar.

Innanfor ordartikkelen kan andre ordformer vera døme på målføreformer av oppslagsordet, avleiingar eller samansetningar, eller dei kan vera (nær)synonym til oppslagsordet eller ord med liknande tyding og bruk. Somme har halvfeit skrift men manglar ordklasse; andre står i kursiv eller rettskrift. Orda kan vera fullskrivne eller stå som ledd i ei rekkje med

sams førsteledd eller etterledd. Ordformer som representerer leksikalske gjenstandar skal registrerast i Aasen-ordbanken, medan bøyingsformer og reine formdøme skal haldast utanfor.

Adel, m. 1) Kjernen el. den bedste deel af Tingen; sædvanlig kun om kjerneved i træ (s. Adelvid). Udtalt: **Adel** og **Ad'el**, Nfj. og fl.**Ad'ul**, Hard.**Ail**, Sdm. og **Al**, mere alm. □ 2) Adel, Herrestand. Nærmest efter det tydske Adel, ligesom de hertil hørende **Adelsmann**, **Adelskap**, **Adelstand**, **adelleg**, adj. og **adla**, v.a. □ 3) i formen Adels, omtr. som: ægte, af rette Slags, f. Ex. **ein Adels Ljugar**: en erkeløgner; **ein Adels fant**: en stor Skalk. Jæd. I danske Dial. adel (ail), Sv. Dial. adal (al): fortrinlig. G.N. adal, i sammensætning, som *adalból*: Hovedgaard; *adalritning*: Hovedskrift.

FIGUR 1. Artikkelen Adel m. i Aasen 1873.

Heterogeniteten i føring gjer fullstendig maskinell uthenting av leksikalske gjenstandar umogleg. Heterogeniteten kjem av at Aasen 1864 og Aasen 1873 er skrivne som laupande tekst der kvar paragraf eller ordbokartikkel skal verka mest mogleg forståeleg, klar og rasjonell. Dermed blir ordninga av element meir mangearta i ordartiklar der eit stort dømetilfang er ordna og drøfta. Eit døme er artikkelen Adel m. (figur 1).Tyding 1 er seinare skild ut som eige ord *al* m. Tyding 2 syner samansetningar og avleiingar. Tyding 3 er no skild ut som eige oppslag. Artikkelen har åtte postar i Aasen-ordbanken. Ord med ordklasse vart registrerte maskinelt.

Båe bøker er svært systematiske, men føreset likevel menneskeleg tolking. Målet er å utføra arbeidet etter Aasens intensjon, som uttrykt i grammatikk og ordartikkel. Seinare tolking, t.d. i Grunnmanuskriptet for Norsk Ordbok, og i sjølve Norsk Ordbok, er konsultert ved behov, men overstyrer ikkje førstehandsopplysningar i Aasen 1864 og Aasen 1873. Om konsekvens i Aasens rettskriving sjå kap. 4.

Uthenting av lemma i artikkeltekst kravde manuelt etterarbeid, gjennomført i to omgangar. I første omgang vart oppslagsformene i fullformgeneratoren kontrollerte mot Aasen 1873, og fekk lagt til paradigme og normeringsstatus for oppslagsord og paradigme.

Under første gjennomgang kom to tilhøve for dagen: (1) Aasen 1873 er organisert litt annleis enn moderne ordbøker i ordning av oppslagsord og definisjonar, noko som har følgjer for tilordning av normeringsstatus, jf. figur 1 og kapittel 6. (2) Ordartiklane har meir informasjon om fleirordsuttrykk og samansetningsledd enn det som før har vore registrert. Etter-

kontrollen er noko over halvvegs fullført (a—l og t—ø), og har ført til om lag 7500 postar i Aasen-ordbanken, slik at summen no ligg på om lag 49 500 lemma.

3. Om paradigme, ordklasse og tvilstilfelle

3.1. Paradigme

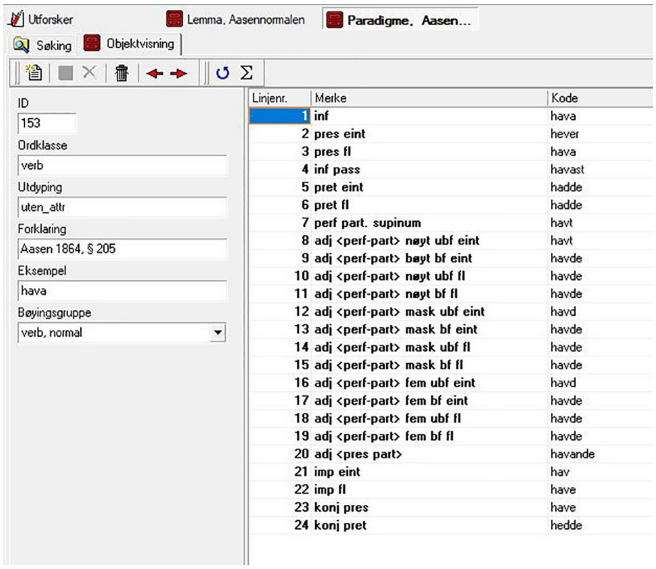
Aasen 1864 har bøyingsskjema for alle store ordklassar, og tabellar over ord med uregelrett (sterk) bøying. Aasens bøyingsskjema for substantiv, adjektiv og verb har fleire kategoriar enn ordbanken for moderne nynorsk. Adjektiv har full bøying i kjønn, tal og binding. Substantiv har kasusformer for dativ og genitiv. For verb er fleirtalsbøying av presens og preteritum og bøying i konjunktiv inkludert. Skjemaet for bøying av fortids partisipp er som for adjektiv, og inkludert i verbskjemaet. Paradigmeskjemaa for Aasen-ordbanken bruker Aasens kategoriar, jf. figur 2 og 3 nedanfor.

	Indikativ.	Konjunkt.	Imperativ.	Infinitiv.	Participium.
Præsens	*hever.*	*have.*	*hav!*	*hava.*	*havd, havt.*
Fleert.	*hava.*	*have.*	*have!*		Fl. *havde*
Imperfekt.	*hadde.*	*(hedde).*			Sup. *havt.*
Fleert.	*hadde.*	*(hedde).*			

Anm. Paa Grund af megen Brug er dette Ord udsat for mange Forkortninger, saasom: ha' (hava), he (hever), hae for 'hadde', som egentlig skulde hedde: havde (G. N. *hafdi*). — Konjunktivet 'have' bruges mest som en Ønskeform, f. Ex. 'Gud have Saali' (ɔ: Sjælen). Det andet Konj. 'hedde' (G. N. *hefdi*) bruges i Hardanger og flere Steder i Formen 'hædde'; f. Ex. 'hædde eg vilja' (ɔ: dersom jeg havde villet). Jf. Tydsk hätte.

FIGUR 2. Verbskjema i Aasen 1864 § 205 for *hava*.

Minuset ved formrikdomen er mange moglege analysar for homografar. Plusset er at skjemaa har kategoriane brukte i registrering av målføre. Aasen 1864 og 1873 dokumenterer norsk talemål ca. 1830—1870, så Aasen-ordbanken kan brukast til å undersøkja talemålsoppskrifter.



FIGUR 3. Paradigmeskjemaet for unikumet *hava* i Aasen-ordbanken.

3.2. Paradigme og ordklasse

Gjennom tilordning av paradigme får kvart oppslagsord ordklasse, jf. tabell 1 nedanfor. Fordelinga samsvarer bra med Ordbanken for nynorsk, men har noko færre paradigme.

Tala omfattar normerte og unormerte oppslagsformer. Talet på oppslagsformer pluss ordklasse er større enn talet på oppslagsformer aleine, fordi nokre ord har meir enn eitt paradigme. Fordelinga mellom ordklassane er som venta, med om lag ein prosent funksjonsord og 98,9 prosent innhaldsord. Substantiv i femininum har større plass i Aasen-ordbanken enn i Ordbanken for nynorsk, jamført med maskulinum og nøytrum (Nynorsk-ordbanken: drygt 123 000 artiklar, 36 prosent m., 18 prosent kvar på f. og n.). Ei nøyaktig kartlegging av genushistorikk i nynorsk finst ikkje, men det er rimeleg å rekna med at både talemålsendring og tilpassing til bokmål i normering av norsk på 1900-talet har spela ei rolle.

TABELL 1. Oppslagsord i Aasen-ordbanken fordelt på ordklasse og med tal paradigme.

Ordklasse	Tal paradigme	Tal lemma	Prosent tal lemma
Substantiv maskulinum	59	12 068	23,39
Substantiv femininum	65	11 235	21,78
Substantiv nøytrum	32	8 321	16,13
Substantiv appellativ (utan genus)	2	20	0,04
Substantiv proprium (eitt paradigme per genus)	3	284	0,55
Adjektiv	73	10 360	20,08
Adverb	3	1 638	3,18
Verb	196	6 885	13,35
Preposisjon	1	298	0,58
Determinativ	37	127	0,25
Pronomen	13	35	0,07
Konjunksjon	1	72	0,14
Interjeksjon	1	53	0,10
Infinitivsmerke	1	3	0,01
Prefiks og førsteledd i samansetningar	1	191	0,37
SUM	488	51 590	100,00

Aasen 1864 tilordnar ordklasse annleis for somme av funksjonsorda enn i moderne grammatikk. Ordklassa determinativ manglar, og overgangen mellom pronomen og adjektiv er sett som uklar (Aasen 1864:173). Funksjonsorda i Aasen-ordbanken har derfor fått same tilordning av ordklasse som Nynorsk-ordbanken bruker for moderne norsk. Dersom Aasen-ordbanken blir brukt i ein morfosyntaktisk taggar, vil det vera ein føremon å ha analysereiskapar som er like nok kvarandre til at resultata kan jamførast med nyare analysar.

3.3. Tvilstilfelle i utforming av paradigme

Adjektivparadigma har vore dei vanskelegaste å koma fram til ei fast løysing for i Aasen-ordbanken. Adjektivbøyinga får etter måten snau omtale i Aasen (1864:158 ff.).

	Eental.			Fleertal.		
	Mask.	Fem.	Neutr.	Mask.	Fem.	Neutr.
Ubestemt F.	*stor* (er).	*stor.*	*stort.*	*store.*	*stora.*	*store.*
Bestemt F.	*store.*	*stora.*	*stora* (e).	*store.*	*store.*	*store.*

FIGUR 4. Bøyingsskjema for adjektiv i Aasen 1864 (s. 158 § 184).

Hovudskjemaet føreset veksling mellom *-a* og *-e* i trykklett utgang i bunden form og fleirtal, også for fleirstava og avleidde adjektiv på *-all*, *-utt*, *-ig*, *-ad* (§ 186, s. 161). Slik er det også framstilt t.d. i Nygaard (1867:20) medan Hægstad ser ut til å opna for både *-e* og *-a* (1879:18-20). Den pragmatiske løysinga kan vera å la dei avleidde adjektiva få to paradigme, eitt normert med standardbøying for adjektiv og eit unormert med gjennomført *-e*.

I gradbøying ser det ut som om superlativ-forma kan enda på *-st*, men til vanleg endar på *-aste* (1864 § 189, s. 165), og skal reknast som ubøyeleg. Her er det likevel opna for båe løysingar, og adjektivparadigma med gradbøying har derfor fått både ubunden form på *-(a)st* og bunden på *-(a)ste*, slik at båe superlativformer skal bli fanga opp dersom dei finst i tekst.

4. Om normeringsstatus i Aasen-ordbanken

Ordbankformatet gjev høve til å merkja oppslagsformer med «normert» eller «unormert». Dermed kan ein skilja oppslagsformer etter Aasen-normalen frå oppslagsformer med ein annan skrivemåte. Å bruka skiljet «normert» versus «unormert» på eit ordregister som må reknast som eit vitskapleg basert framlegg om mønsterformer, kan verka som ei tvangstrøye påført Aasen 1873 i ettertid; men Aasen sjølv var svært oppteken av at rettskrivinga for landsmålet burde vera regelfast, tilnærma unntakslaus og med få homografar (Aasen 1957 II:105–7). I Aasen-ordbanken er Aasens vurderingar og intensjonar lagde til grunn så langt råd.

Skiljet mellom «normert» og «unormert» for oppslagsord og paradigme kan brukast til å nyansera utforskinga av tidleg nynorsk tekst. Om ein ser etter samsvar først med «normert», deretter «unormert», vil resultatet seia noko om heterogenitet versus einskaplegheit i den tekstmengda ein undersøkjer.

4.1. Normeringsstatus for oppslagsform

Avgjerda om å merkja ei ordform som «normert» eller «unormert» byggjer på både ordform og kontekst. Aasen 1873 inneheld mange artiklar med tilvisingsformat, men tilvisingar gjeld ikkje alltid form. Tommelfingerregelen i registreringa vart at tilvisingsartiklar av typen 1–2 nedanfor truleg kunne setjast som unormerte, men ein må alltid kontrollera mot målartikkelen.

(1) **abakleg**, s. avbakleg. (målføreform — status «unormert»)

(2) **andleg**, s. andeleg. (alternativ oppslagsform for ordartikkel — status «normert»)

Døme 3-4 viser artiklar som er tilvisingsartiklar i forma, men ordet det blir vist til, kan ikkje vera ei form av oppslagsordet.

(3) **Allhelgamess**, s. Helgamessa. (synonym, i målføreform under målartikkelen – oppslagsord status «normert»)

(4) **Augnevar**, s. Augnesaur. (synonym — oppslagsord status «normert»)

Dersom ein artikkel inneheld meir enn ei rein tilvising, til dømes ei tydingsopplysning, kan ein normalt rekna med at oppslagsordet skal ha status normert, men det må kontrollerast.

Sidan tilvisingspraksis i Aasen 1873 er fleirtydig, er normeringsstatus vurdert for kvar oppslagsform. Unormerte oppslagsord er annoterte i kommentardelen av fullformgeneratoren.

4.2. Normeringsstatus for oppslagsform kombinert med normeringsstatus for paradigme

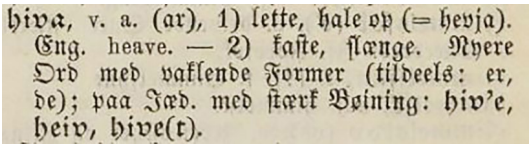
Aasen-ordbanken krev at ei normert oppslagform må ha minst eitt normert paradigme. Tilleggsparadigme kan vera normerte eller unormerte. Det er eit mål å finna paradigme innanfor Aasen-normalen for alle oppslagsformer, men det er òg laga nokre paradigme som berre blir nytta for unormerte oppslagsformer, til dømes for svake femininum på *-e*.

Unormerte oppslagsformer kan ha eitt eller fleire unormerte paradigme, og kan mangla paradigme. Dette gjev grupperinga i tabell 2, med tal tilslag for gruppene.

TABELL 2. Tilslag i Aasen-ordbanken for normeringsstatus av oppslagsform med paradigme.

Gruppe	**Tal tilslag**
Normert oppslag	44545
1 Normert oppslag, normert paradigme	44051
2 Normert oppslag, unormert paradigme	494
Unormert oppslag	5756
1 Unormert oppslag, unormert paradigme	4758
2 Unormert oppslag, manglar paradigme	998

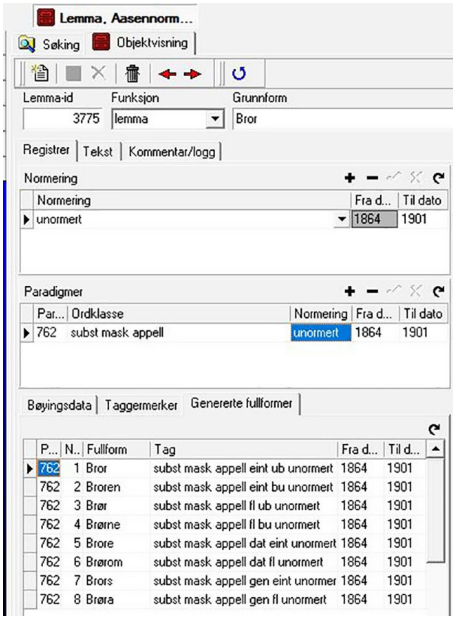
Aasen 1873 har om lag 49500 oppslagsformer i alt. Tabell 2 syner at ei normert oppslagsform normalt har eitt paradigme. Dei unormerte paradigma for normerte oppslagsformer er alle tilleggsparadigme.


hiva, v. a. (ar), 1) lette, hale op (= hevja). Eng. heave. — 2) kaste, slænge. Nyere Ord med vaklende Former (tildeels: er, de); paa Fæd. med stærk Bøining: hiv'e, heiv, hive(t).

FIGUR 5. Aasen 1873. Verbet *hiva* v.a. med tre bøyingsalternativ

Eit døme er verbet *hiva* (figur 5), som er sett opp med a-bøying som hovudform, og bøying med preteritum *-de* og med sterk bøying som alternativ. Desse er registrerte i Aasen-ordbanken som unormerte. I moderne nynorsk har *hiva* sterk bøying eller *-de* i preteritum, medan a-bøyinga er gått ut or rettskrivinga.

I Aasen-ordbanken kan unormerte oppslagsformer ha unormerte paradigme eller mangla paradigme. Dersom oppslagsforma høver med krava til grunnform i Aasen-normalen, er det føydd til paradigme. Eit døme er *Bror* m., nemnt som vanleg uttaleform i ordartikkelen *Broder* m. I 1901 kom forma *bror* m. inn i rettskrivinga, og var før det karakterisert som den vanlege skriftforma (Hægstad et al. 1899:42). Denne fråsegna kan ein testa via Aasen-ordbanken.



FIGUR 6. Aasen-ordbanken. Substantivet *Bror* m., status og paradigme unormert.

Opp mot 1000 unormerte oppslagsformer kan ikkje få paradigme av di dei har ei form som ikkje høver i Aasen-normalen, men dei er registrerte som oppslagsformer i Aasen-ordbanken. Mange er jamningsformer, t.d. *aamaa*, *Bakvudu*, *Broto*, *Hugu*. Andre er bøygde former frå paradigme for andre oppslag (*eikor*, *eitkvart*, *flaut*, *rauk*) eller av substantiv, som *Klædevegen*, *Trinsekakur*. I tekst vil dei bli attkjende som bøyingsformer for ei normert grunnform med paradigme registrert i Aasen-ordbanken: *Klædeveg* m, *Trinsekaka* f.

5. Funn av oppslagsformer ved manuell gjennomgang

Den manuelle ekserperinga har så langt ført til 7000 nye oppslagsord-postar i Aasen-ordbanken. Av desse har om lag 500 meir enn eitt paradigme, eller dei er kopla til fleire artiklar.

TABELL 3. Ordklassefordeling for manuelt eksperpert tilfang frå Aasen 1873.

Ordklasse	**Tal oppslagsord**	**Tal oppslagsord + paradigme**
Utan ordklasse	143	143
Substantiv maskulinum	1750	1830
Substantiv femininum	1253	1363
Substantiv nøytrum	1102	1177
Substantiv appellativ	6	6
Substantiv proprium	209	225
Adjektiv	1279	1323
Adverb	490	512
Verb	394	426
Preposisjon	37	123
Determinativ	17	19
Pronomen	12	12
Konjunksjon	16	17
Interjeksjon	23	19
Infinitivsmerke	2	2
Prefiks og førsteledd i samansetningar	161	172

Ekserperinga er ikkje over, om lag 60 prosent av ordboksteksten er finlesen. Det ser ut til at om lag 20 prosent oppslagsorda i Aasen-ordbanken kjem frå artikkeltekst.

Somme leksikalske gjenstandar er registrerte med både målføreform og normert form. Aasen 1873 kan ha målføreform på oppslagsplass i ein tilvisingsartikkel, og normert form som overskrift på hovudartikkelen. Oppslagsord i artikkeltekst har oftast normert oppslagsform.

Før 1900 var bruken av bindestrek som leddelingsmerke ikkje ortografisk regulert. Dette er løyst ved at dei om lag 1500 oppslagsformene med bindestrek i Aasen 1873 har ei tilsvarande form utan bindestrek i Aasen-ordbanken.

5.1. Samansetningar og avleiingar

Om lag 80 prosent av det tilleggsekserperte tilfanget er substantiv eller adjektiv. Det meste er samansetningar som viser fram ordlagingspotensialet i oppslagsordet. Eit døme er *burt* adv.

> Ordet træder i Forbindelse med mangfoldige Participier, f. Ex. **burtblaasen** (bortblæst); saaledes med: boren, dregen, fallen, faren, flutt, gjeven, havd (dvs. bragt), kallad, kastad, komen, lagd, laten, leigd, lovad, maadd, reken, rømd, seld, sett, slegen, stolen, teken, vend, vikt, og lignende. (Aasen 1873:91)

Denne ordrekkja er ikkje framheva typografisk. Ho er der for å syna ordlagingsfunksjonen til adverbet *burt*. Somme av døma er ikkje redigerte i Norsk Ordbok. Alle krev sjølvsagt tilleggstilfang, men først må ein vita om dei. Det vil ein få støtte til gjennom Aasen-ordbanken.

5.2. Fleirordsuttrykk

Tilleggsekserperinga har registrert om lag 300 fleirordsuttrykk. Om lag 220 av desse er adverb av typen «preposisjon + substantiv», med substantiv i dativ eller genitiv. Slike ordlag har tradisjonelt vore handsama under hovudordet i ordbøker, men i rettskrivingsordlister må dei stå på alfabetisk plass. Mange er svært vanlege, t.d. *i Aftan*, *just som*, *um Bord*. I Aasen 1873 kan slike ordlag stå under kvart ord i ordlaget, sjeldan i eigne artiklar.

Det er interessant at Aasen 1873 har desse uttrykka særskrivne, for i samtidsdansken var samanskriving vanleg, medan særskriving er regelen i norsk no. Men det viktigaste er at desse uttrykka er registrerte som leksikalske gjenstandar med tilknytte omgrep uttrykte i definisjonar.

Ei anna gruppe fleirordsuttrykk er substantiviske, og på veg til å bli samansetningar. Døme er *Vetters Dag*, *Litle Jolaftan*, *Klædes Trøya*. Desse er registrerte med paradigmet for etterleddet.

5.3. Førsteledd i avleiingar og samansetningar

I samansetningsspråk som norsk har all tekst sjeldsynte samansetningar som må leddelast for å tolkast rett. Aasen 1873 har nokre artiklar for avleiingsprefiks og avleiingssuffiks. Desse står utan ordklasse eller markering av at leddet er usjølvstendig, men dei første orda i definisjonen er oftast «en Partikkel som …». I Aasen-ordbanken har slike ord eige paradigme (nr. 1002), med den forma som Aasen 1873 fører opp.

I den manuelle gjennomgangen er alle førsteledd i samansetningar registrerte dersom dei skil seg i form frå oppslagsordet, t.d. *Armods-* (av

Armod), *Augne-* (av *Auga*), *Ferda-* (av *Ferd*). Er det fleire ulike førsteledd, blir alle registrerte, til dømes *Dreng-* og *Drengje-* (av *Dreng*).

Aasen 1864 poengterer at genitiv blir nytta «kun i den ubestemte Form og mest i Sammensætning» (s. 135). Det er ofte samanfall mellom førsteleddet og genitivsforma til grunnordet, men ikkje i slik grad at ein kan la vera å registrera førsteledd.

5.4. Namn

Aasen 1873 har artiklar for namneledd, med namn som døme, jamfør *mund* m., *veig* f., for personnamn med ordklasse etter kjønn. Dersom eit appellativ kan brukast i namn, er dette handsama som ei vanleg tydingsopplysning. Artikkeloverskriftene er registrerte maskinelt. I etterraksten er også døma på namn tekne med i Aasen-ordbanken, registrerte som proprium.

I maskinanalyse av tekst er attkjenning av namn notorisk vanskeleg. Det gjeld særleg namneformer som ortografisk fell saman med appellativ. Namneinventaret i Aasen 1873 er lite jamført med det ein finn i tekstanalyse, men det er ei byrjing. For studiet av Aasen 1873 har det interesse å sjå korleis proprium er handsama, opp mot appellativ.

6. Artikkeltypar, indre samanheng og etymologi

I moderne ordbøker er det til vanleg klare skilje mellom artikkeltypar. Hovudskiljet går mellom innhaldsartiklar, som fortel om oppslagsordet, og tilvisingsartiklar, som peikar frå ei ordform til ei anna. Det er underforstått at oppslagsforma for tilvisingsartikkelen høyrer saman med oppslagsforma i målartikkelen, og vil bli attfunnen der.

Aasens normalartiklar har faste kategoriar, men ordartiklane er mindre skjematisk oppbygde enn i ei moderne ordbok. Dei fastaste felta i ordartiklane er artikkeloverskrifta og det som følgjer rett etter: oppslagsform, oftast i grunnform, ordklasse og bøyingsopplysningar. I somme tilfelle er ei anna ortografisk form på oppslagsordet med, likeins uttaleopplysning. Definisjonen følgjer oftast rett etter. Deretter er det stor innhaldsvariasjon.

6.1. Tilvisingsartikkel i form

Svært korte artiklar er ofte tilvisingsartiklar, men tilvisinga har fleire funksjonar enn å peika på formvariasjon. Aasen 1873 bruker tilvisingssystemet til å kasta lys over tydingsslektskap. Fleirordsuttrykk kan til dømes ha sjølvstendige men snaue artiklar:

(5) **aa gange**: paa Færde; s. Gang.

Under *Gang* m. finn ein variantar, komparentformer og utdjupande kommentarar. Døme 5 syner kva uttrykket tyder og peikar mot artikkelen der hovudordet er handsama. Redigeringsmåten kan vera vald fordi han stør eit hovudføremål med Aasen 1873: å syna den indre samanhengen i norsk talemål.

Døme 6 er ei rein innhaldstilvising til døme 7:

(6) **Sorphøna**, f. et Legetøi, s. Snørekall.

(7) **Snørekall**, m. et Slags Legetøi; en Pind, som er stukken igjennem en rund Skive og saaledes afpasset, at den kan sættes i en hvirvlende Bevægelse, ligesom en Haandteen. … Ogsaa kaldet Snørebasse, el. Snurrbasse …, Snørebuss …, Sorphøna …, Gand.

Krysstilvisingar blir brukt i Aasen 1873 meir som i leksikon enn i semasiologiske ordbøker.

6.2. Oppslagsord grupperte rundt ein definisjon

Aasen 1873 har mange døme på at oppslagsordet opnar for å seia noko om ei gruppe synonyme eller nærsynonyme ord med ulik form. Fleire formlar peikar mot synonymi og andre slag tydingssamband, såleis teiknet «=» (3227 førekomstar, tyder ei eller anna form for ekvivalens), «også kaldet» (372 førekomstar). Resultatet kan vera ein artikkel sentrert rundt definisjonen, som i døme 8. Forma med enklast struktur er oppslagsform. To av oppslagsformene som følgjer, har tilvisingsartikkel til denne artikkelen, dei andre to finst berre her.

(8) **Kjøta**, f. Kjødside, Indside paa Skind eller Huder. Hard. Ogsaa kaldet: **Kjøkka** (Kjøtka), Hall., **Kjøtska**, Buskr. Ellers: **Kjøtrosa** (o'), f. B. Stift, Nordl. **Kjøtroslid** (i'), f. Sæt.

Formelen «hedder ogsaa» (464 førekomstar) kjem i forlenging av ein definisjon som innleiing til nærskylde avleiingar eller samansetningar førte som synonymdøme. Desse har somtid eigen artikkel, somtid ikkje. I døme 9 har *naudkyta* eigen artikkel, *naudskrala* finst berre her.

(9) **naudskreppa**, v.n. skryde idelig el. overmaade. Hedder ogsaa **naudkyta** (naukjyte) og **naudskrala**. (Hall.).

Slik organisering av innhald bryt med det ein ventar av ei semasiologisk ordbok. Det er velkjent at Aasen var oppteken av onomasiologi. Det ser ein av Aasen 1864 (Fjerde Afdeling IV Orddannelse efter Betydningen), og av tesaurusordboka *Norsk Maalbunad*. Dess viktigare blir det å få ei nøyaktig registrering av alle leksikalske gjenstandar som er tekne med i Aasen 1873.

7. Konklusjon

Ivar Aasen ville at Noreg skulle ha eit eige skriftspråk som skulle stetta alle dei krav som hans samtid sette til standardspråk. *Norsk Ordbog med dansk Forklaring* må sjåast som ei vitskapleg ordbok, der føremålet er å leggja fram all tilgjengeleg og pålitande informasjon om norsk talespråk på ein slik måte at sjølve språket får legitimitet, både i høve til eksistens og status, ved at skriftnormalen kan brukast til alt som eit utvikla skriftspråk blir brukt til (Hoel 2018:28).

Målet vart nådd med Aasen 1873, ved at ordboka vart vørd som eit storverk, knapt møtte kritikk og straks vart teken i bruk. Men meldingane frå vitskapsmiljøet var få og snaue, og noka drøfting av materialbruk eller metode i Aasen 1873 kom ikkje. Det skal ha vore eit vonbrot.

Rettskrivinga i Aasen 1873 vart teken i bruk, men modifisert i 1901. Ordtilfanget i Aasen 1873 er framleis sentralt i norsk, og mykje finst i standardordbøker i uendra form (Ore & Grønvik 2018b). Korleis det nynorske ordtilfanget arta seg før og etter at Aasen 1873 kom i bruk, får vi sjå når større mengder av tidleg nynorsk tekst blir undersøkte med Aasen-ordbanken.

Referansar

Ordbøker

Grunnmanuskriptet for Norsk Ordbok 1940. Manuskript. Dokumentasjonsprosjektet. <https://usd.uib.no/perl/search/search.cgi?tabid=993&appid=59> . Henta mars 2023.

Norsk ordbank – nynorsk 2012 2018. <https://www.nb.no/sprakbanken/ressurskatalog/oai-nb-no-sbr-41/>. Henta mars 2023.

Norsk Ordbok. Ordbok over det norske folkemålet og det nynorske skriftmålet. I–XII. 1950–2016. Hovudredaktørar: A. Hellevik, L. Vikør, O. Grønvik, L. Killingbergtrø, D. Worren & H. Gundersen. Oslo. Samlaget.

Aasen, Ivar 1873. *Norsk Ordbog med dansk Forklaring*. Christiania. Mallings Boghandel.

Aasen, Ivar 1925. *Norsk maalbunad. Samanstilling av norske ord etter umgrip og tyding*. Oslo. Samlaget. Utg. v. Sigurd Kolsrud.

Aasen-ordbanken. Lemma 2021<https://usd.uib.no/perl/search/search.cgi?appid=250&tabid=3557>. Henta april 2023.

Annan litteratur

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

Hoel, Oddmund 2018. Norsk språkhistorieskriving. I: Sandøy, Helge & al. (red.) 2018, *Norsk språkhistorie. III : Ideologi*. Oslo. Novus, 51–148.

Hægstad, Marius 1879. *Norsk Maallæra elder Grammatik i Landsmaalet*. Eiget Forl.

Hægstad, Marius, Rasmus Flo & Arne Garborg 1899. *Framlegg til skrivereglar for landsmaale i skularne*. Christiania. Brøgger.

Kruken, Kristoffer & Terje Aarset 2003. Innleiing. I: Ivar Aasen, *Norsk Ordbog med dansk Forklaring*. Ny utgåve ved Kruken og Aarset. Oslo. Samlaget, IX—XXXVI.

Nygaard, Marius 1867. *Kortfattet Fremstilling af det norske Landsmaals Grammatik*. Bergen. Gjertsen.

Ore, Christian-Emil Smith & Oddrun Grønvik 2018a. Bokmål og nynorsk samindeksert – Metaordboka som verktøy for jamføring og

utforsking av ordtilfang. I: Svavarsdóttir, Ásta, Halldóra Jónsdóttir, Helga Hilmisdóttir & Þórdís Úlfarsdóttir (red.), *Nordiske Studier i Leksikografi 14*. Reykjavík: Nordisk Forening for Leksikografi, 87–95.

Ore, Christian-Emil Smith & Oddrun Grønvik 2018b. Comparing Orthographies in Space and Time through Lexicographic Resources. I: Čibej, Jaka, Vojko Gorjanc, Iztok Kosem & Simon Krek (Eds.), *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*. Znanstvena založba Filozofske fakultete, 159–172.

Ore, Christian-Emil Smith, Oddrun Grønvik & Trond Minde 2023. Et fullformsystem for analyse av eldre tekst på tidlig nynorsk, bygd på Aasen-normalen. I: Holmer, Louise et al. (red.), *Nordiska studier i lexikografi 16*. Lund & Göteborg: Nordiska föreningen för lexikografi, 267–279.

Aasen, Ivar 1864. *Norsk Grammatik*. Christiania: Mallings Bogtrykkeri.

Aasen, Ivar 1957. *Brev og dagbøker 1—3*. 1. Brev 1828–1861. 2. Brev 1862—1896. 3. Dagbøker 1830—1896. [red. Reidar Djupedal]. Oslo. Samlaget.

Dansk-svensk/svensk-dansk onlinebaseret flerfagsordbog

Poul Hansen

This article presents a new Danish-Swedish/Swedish-Danish thematic dictionary for professional translators, writers, and communicators. After a review of 1) existing specialist dictionaries between Danish and Swedish, 2) the most demanded subject areas on the translation market, and 3) a description of specialist translation work in practice, the most important characteristics of the new dictionary are described. The dictionary contains 160,000 entries in both language directions and describes 60 different themes that are relevant in connection with professional translation between Danish and Swedish and post-editing work.

KEYWORDS: flerfagsordbog, fagleksikografi, oversættelse, efterredigering, dansk/svensk

1. Indledning

Oversættere lever i dag i en tidsalder, hvor information spiller en stadigt større rolle. Og det er en virkelighed, hvor der ofte stilles meget høje krav til detaljer og præcise beskrivelser. Søgning efter adækvat information er vigtigere end nogensinde før, i særdeleshed for den, der arbejder professionelt med tekstproduktion. Desuden produceres der i dag flere fagtekster end tidligere. Uden redskaber til at navigere gennem strømmen af mere eller mindre korrekte informationer er det svært at producere fagoversættelser på et højt professionelt plan. Her spiller fagleksikografi en vigtig rolle med hensyn til at bistå fagoversætteren, og fagleksikografiens udfordringer og muligheder i den forbindelse er blevet diskuteret indgående i litteraturen, se f.eks. Bergenholtz & Tarp (1994), Fuertes-Olivera & Tarp (2014) og Leroyer (2018). Der er også søsat flere fagleksikografiske projekter med baggrund i disse overvejelser, f.eks. Ejendomsordbogen fransk/dansk (Leroyer & Kruse 2012) og Vinsmagningsordbogen Œnolex Bourgogne (Leroyer & Høy 2016).

Den flerfagsordbog, som præsenteres her, skriver sig også ind i denne fagleksikografiske kontekst. Ordbogen blev publiceret i februar 2021 i den netbaserede ordbogssamling WordFinder Unlimited under navnet *Dansk-svensk*

ordbok, og den indeholder omkring 160.000 opslagsord i begge sprogretninger inden for en række af de områder, hvor der i dag efterspørges oversættelse mellem dansk og svensk. Ordbogen er udviklet af to erfarne translatører, som driver firmaet Öresunds Översättningsbyrå, og er primært tænkt som et praktisk værktøj til informationssøgning i forbindelse med oversættelses- og efterredigeringsarbejde, men er ikke begrænset til dette. Den kan også bruges til at kontrollere og supplere information fra andre kilder. Målgruppen er fagoversættere, tolke, technical writers, copy editors, studerende inden for fagligt orienterede områder, journalister, bibliotekarer og i det hele taget praktisk virksomme akademikere, fagfolk og engagerede lægfolk.

Udviklingen af ordbogen er sket på baggrund af, at al finansieret udvikling af terminologi- og fagsprogsresurser i Norden i princippet er ophørt, hvad der i høj grad påvirker arbejdssituationen for dem, der oversætter professionelt mellem dansk og svensk. I 2001 blev Rådet for teknisk terminologi (RTT) i Norge nedlagt efter i 63 år at have udgivet en lang række flersprogede fagordbøger, blandt andet med dansk og svensk. I 2018 blev Tekniska nomenklaturcentralen (TNC) i Sverige nedlagt efter i 77 år at have udgivet et stort antal specialordbøger, hvoraf flere medtager dansk som et af sprogene. Nedlæggelserne blev begrundet med manglende økonomi og skyldes muligvis også en forventning om, at den hastige udvikling inden for maskinoversættelse, AI og sprogteknologi skulle gøre deres rolle overflødig.

I dag er der brug for en større anerkendelse af, at der fortsat er behov for adækvat oversættelse og specialiserede ordbogsresurser mellem dansk og svensk. Sådanne ordbogsresurser vil med fordel kunne inddrages i sprogteknologi (f.eks. databaser integreret i CAT-værktøjer, oversættelseshukommelser samt træning af AI-systemer).

Det bredere formål med ordbogen er naturligvis også at gøre noget ved det problem, at mange mennesker, særligt i faglige sammenhæng, ikke er bevidst om de specielle faldgruber, der findes med oversættelse mellem dansk og svensk, og dermed kan være alt for ukritiske over for den information fra internettet, der er oversat automatisk. Det, der umiddelbart føles overbevisende korrekt, kan være helt forkert. Oversættelsesprogrammerne kan også producere tekster, som tilsyneladende kan virke helt forsvarlige set ud fra et almensprogligt perspektiv, men som er forkerte set i forhold til kildetekstens fagområde.

Et lille eksempel på denne problematik er danske og svenske plantenavne inden for havebrug, hvor der forekommer lumske inkonsekven-

ser, hvilket tydeligt kan ses på frøposer, f.eks. at plantearten *Echinops ritro* kan hedde både *kugletidsel* og *tidselkugle* på dansk, mens kun navnet *bolltistel* bruges på svensk. Et automatisk oversættelsesprogram vil oversætte – i hvert tilfælde indtil nu – *tidselkugle* til det fejlagtige *tistelboll* (i stedet for det korrekte *bolltistel*).

Leksikografi, indeksering, katalogisering og resurseudvikling er således blevet vigtigere end nogensinde tidligere som et supplement til den information, man finder online.

2. Tidligere udgivne fagterminologiske resurser mellem dansk og svensk

Der er relativt få eksempler på fagterminologiske ordbøger, som kun beskæftiger sig med dansk og svensk. Et eksempel er den i dag stærkt forældede *Teknisk svensk-dansk ordbog* (Linder 1965) med ca. 33.000 opslagsord, og et andet er *Svensk-dansk fackordbok* (Orlando 2006) med ca. 23.500 opslagsord, bl.a. en del tekniske ord. I øvrigt handler det mest om nordiske flersprogede ordlister eller termlister udgivet med henblik på at opnå en klar og ensartet teknisk terminologi. De største resurser inden for oversættelse mellem dansk og svensk fagsprog findes i form af multilingvale samlinger eller databaser, f.eks. Rikstermbanken og IATE (Interactive Terminology for Europe), som er EU's terminologidatabase med millioner af termer, der trods det imponerende omfang har en ufuldstændig fagområdedækning set i forhold til det faktiske behov inden for oversættelse mellem dansk og svensk.

3. Fagområder, der i dag er dækket af specialordbøger mellem dansk og svensk

I Pálfi (2011) gennemgås de fagområder, hvor der i 2011 forelå specialordbøger mellem dansk og svensk i bogform eller online, se Tabel 1, men sammenlignet med andre sprogpar, er udbuddet af fagordbøger meget begrænset. Der mangler desuden mange områder, som er udfordrende for fagoversættere mellem dansk og svensk, f.eks. stillingsbetegnelser i erhvervslivet, havebrug, termer inden for socialforvaltning og offentlig forvaltning, filmtitler, kognitiv adfærdsterapi og mindfulness, bare for at nævne nogle eksempler. Mange af de nedennævnte specialordbøger karakteriseres også af at være korte eller ufuldstændige. Det kan også tilføjes,

at mens der er udarbejdet omfattende specialordlister for tolke mellem svensk og flere fremmede sprog, mangler tilsvarende for dansk og svensk.

TABEL 1. Oversigt over specialordbøger mellem dansk og svensk i 2011.

Algoritmik	Fugle*	Maskinteknik	Statsnavne
Arbejdskøretøjer*	Gastronomi	Militærvæsen	Strikning
Arbejdsmarked	Geologi	Miljø*	Svampe*
Arkivarbejde	Industri	Nye ord*	Svejseteknik*
Bandeord/skælds-	Informationsteknologi	Plantesygdomme	Søfart
ord*	Insekter	Plasmaskæring*	Teater
BDI	Jura*	Plastindustri	Tekstildesign
Bilteknik*	Kemi	Pædagogik	Træsorter
Bjergværk	Klædedragt/stof*	Samfundsvidenskab	Transport
Bliss	Knuder	Skat, moms og andre	Transportinformatik
Boligmarked	Kontormateriale*	udgifter	Typografi
Botanik*	Korrosion	Skånsk	Uddannelse
Byggeteknik*	Landbrug*	Slang*	Ungdomssprog*
Cykling	Leksikografi	SMS- og chatsprog*	Vejteknik
Etymologi	Levnedsmiddel-	Socialforvaltning*	Zoologi
Fisk	videnskab*	Spejdersprog	Økonomi
Fiskeri	Lim	Sportsfiskeri*	
Forvaltning	Luftfart	Statistik	

Asterisk "*" markerer de fagområder, hvor Pálfi (2011) refererer og angiver direkte links til Öresunds Översättningsbyrås separat publicerede online-ordlister.

4. Fagoversætterens brugerperspektiv

I fagordbøgers forord kan man ofte læse, at ordbogsproducenterne prøver at tage udgangspunkt i brugernes behov, arbejdssituation og forudsætninger, hvilket naturligvis er meget prisværdigt, eftersom ordbøgerne jo er til for at hjælpe deres intenderede brugere. Men trods de gode intentioner er det som regel svært for ordbogsproducenterne at leve op til deres erklærede hensigt, i hvert fald set ud fra en oversætters synsvinkel, og det er sjældent, at oversættere bliver spurgt og får lov til at påvirke ordbogsarbejdet. Et andet problem er, at det mange gange kan konstateres, at angi-

velse af korrekte ækvivalenser og oplysninger om deres brug ofte må vige for en overvejende faglingvistisk holdning, bl.a. med den konsekvens, at ordbogen hellere forklarer et vanskeligt oversætteligt ord end foreslår en anvendelig praktisk og situationsnær oversættelse. For oversættere er det vigtigste jo, at de kan få et rimeligt og hurtigt svar på et specifikt oversættelsesproblem, særligt i form af ækvivalenter. Upræcise oversættelsesløsninger og lang søgetid efter alternative løsninger kan være et irritationsmoment for en fagoversætter, fordi det medfører uønsket tidsforsinkelse. Til sidst kan det ofte konstateres, at udvalget af fagord i almensproglige ordbøger håndteres og vurderes alene i forhold til ordenes gennemslag i almensproget. Det svarer ikke nødvendigvis til den virkelighed, som en fagoversætter lever i, hvorfor dette næppe kan være til gavn for fagoversætteren.

5. Oversættelse af dansk og svensk fagsprog i praksis

For den, som har været med siden halvfjerdserne, hvor man skrev på skrivemaskine og leverede oversættelser på maskinskrevne papirark, har udviklingen været enorm. Nye hjælpemidler har øget produktionshastigheden betydeligt i forhold til tidligere. I dag er det ikke ualmindeligt, at oversætterens praktiske arbejde er reduceret til at efterredigere og bearbejde tekster, som er genereret af et oversættelsesprogram. De moderne hjælpemidler og arbejdsmetoder løser dog langt fra de grundlæggende problemer, som en dansk-svensk fagoversætter kæmpede med for 30 år siden og stadig kæmper med i dag. Der mangler stadig tilstrækkeligt omfattende ordbøger, og der er stadig alt for ofte behov for at udnytte et tredje sprog som mellemled i jagten på den korrekte term. Takket være internettet er det dog nu muligt at bruge billeder og finde referencetekster i en takt og en udstrækning, som tidligere ikke var mulig. Dette er dog et tidskrævende arbejde. Og uden oversætterens egen sproglige formåen og gode terminologiske resurser at slå op i, kan der ikke leveres en god oversættelseskvalitet.

I virkeligheden er den afsluttende kvalitetssikring af en oversættelse blevet langt mere tidskrævende og udfordrende end tidligere. Man skal være ekstra opmærksom på en ny og fagligt set mere intellektuelt krævende måde, og man skal foretage mange ekstra opslag, når oversættelsen indeholder termer, man ikke selv ville have valgt, men som et (gratis) oversættelsesværktøj har foreslået. Som oversætter kan man meget, men ikke alt,

og hver oversætter har desuden sin egen stil og måde at udtrykke sig på. Det gælder særligt i efterredigeringssituationer.

6. Markedets efterspørgsel efter fagoversættelse mellem dansk og svensk

Mange fagtekster er i dag på engelsk, og det er ikke usædvanligt, at aktører på markedet spontant giver udtryk for, at det ikke er nødvendigt at oversætte tekster mellem dansk og svensk, når man har dem på engelsk. Men der er stadig mange områder, hvor engelsk er helt uegnet, og hvor det er af afgørende vigtighed, at tekster foreligger i både en dansk og en svensk version. Det er vanskeligt at forestille sig, at behovet for professionel oversættelse mellem dansk og svensk helt skal ophøre.

Efterspørgslen efter oversættelse af fagtekster mellem dansk og svensk hænger nært sammen med den igangværende udvikling i samfundet, især inden for teknologi, hvor nye fagord hele tiden introduceres og skaber behov for oversættelse. Da cd-pladen blev udviklet, måtte oversætterne bl.a. forholde sig til, hvordan man skulle betegne overfladen på cd'en. Skulle man skrive, at den var *iriserende* eller *regnbuefarvet*? Ingen daværende ordbøger kunne hjælpe der. I dag er lignende problemer opstået i forbindelse med opdelingen af affald i fraktioner. For oversættere har det indebåret, at man skal tage stilling til mange nye begreber i forbindelse med affaldsindsamling og sortering, hele vejen fra udformning af affaldscontainere og beskrivelse af renovationsbilers nye sofistikerede tekniske indretninger til kommuners udbudsdokumenter vedrørende indkøb af affaldsservice og nye renovationsbiler. Behovet for oversættelse mellem dansk og svensk opstår som en følge af, at udbydere fra hele Norden er engagerede i sådanne forandringsprocesser. Det betyder, at fagordbøger hele tiden skal opdateres for at kunne give hjælp til oversættelse af nyskrevne fagtekster, ofte med nye termer. Ikke sjældent drejer det sig om ting, der befinder sig i lanceringsfasen eller endnu ikke er ude på markedet.

I Tabel 2 præsenteres, med udgangspunkt i forespørgsler til Öresunds Översättningsbyrå i årene 1987-2022, en oversigt over nogle af de typiske fagområder, hvor der er et marked for oversættelse mellem dansk og svensk.

TABEL 2. Fagoversættelse ved Öresunds Översättningsbyrå 1987-2022.

- Alternativ medicin
- Annoncer til aviser og blade
- Banksager, prospekter, finansiel information, konkurser
- Byggeri, arkitektur, indretning, byggematerialer
- Dyrkning og plantepleje, gødningsprodukter, bekæmpelsesmidler
- Dåbsattester, vielsesattester, testamenter
- Forretningsaftaler, vedtægter, forsikringsvilkår
- Hjemmesider
- Husholdningsprodukter, personlig pleje, beklædningsprodukter
- GDPR, virksomhedsinterne dokumenter, personalehåndbøger
- Jargon og talesprog, transskribering/oversættelse af indspillede forhør, telefonaflytninger
- Jobansøgninger, CV
- Kriminalitet, politiarbejde, retsprocesser
- Landbrugsdrift, maskiner
- Madlavning, opskrifter
- Medicin og sundhed, patientkontakt, behandlingsmetoder
- Medicinsk forskning og udvikling, dokumenter til Det Etiske Råd, specialprodukter, kliniske forsøg
- Messer, udstillinger, arrangementer, kampagner
- Miljø, økologi, biologi, kemi, kemikaliebrug
- Produktbeskrivelser, emballagetekster, varedeklarationer, advarselstekster
- Social sagsbehandling, klientkommunikation
- Softwareapplikationer, brugerinterface, indbyggede orddatabaser
- Spørgeskemaer, blanketter, undersøgelser, analyseskemaer
- Teknik med høj detaljeringsgrad, manualer
- Turisme, brochurer
- Undertekstning af informations-, undervisnings- og salgsvideoer, speakertekster
- Undervisning og forskning
- Virksomhedsetableringer i nabolandet
- Værktøj

Hertil kommer områder inden for hobby- og fritidsområdet, dvs. områder, hvor mennesker forgæves har søgt information på nettet og somme tider henvender sig til fagoversættere i håb om at få gratis hjælp med enkelte ord, f.eks. forældede ord (slægtsforskere), håndarbejdsord og sjældne forkortelser.

Det sidstnævnte kan også ses som et eksempel på, at flerfagsordbøger er vanskelige at gøre helt dækkende. Nærværende flerfagsordbog kan kun delvis imødekomme almenhedens behov for oversættelse af termer af ikkekommerciel karakter, og der er fagområder, hvor enkeltfagsfagordbøger har bedre forudsætninger for at dække behovet – en problematik, som også er blevet diskuteret i litteraturen (Bergenholtz & Tarp 1994).

7. Historikken bag tilblivelsen af ordbogen

Initiativet til en ordbog mellem dansk og svensk for oversættere blev taget i 2005 på baggrund af en række internt udviklede ordlister i forbindelse med dansk-svenske oversættelsesprojekter siden 1988. Dengang blev det mere almindeligt blandt oversættere at opbygge egne multilingvale termlister, som kunne udgøre en funktionel bestanddel i forskellige oversættelsesværktøjer. Der opstod internationale oversætterplatforme (f.eks. Proz.com), som sammenstillede termlister, og der fandtes centrale diskussionsfora mellem nordiske oversættere, hvor man hjalp hinanden med termoversættelse. Men den praktiske betydning af sådanne individuelle satsninger og deling af resurser er mindsket, bl.a. fordi mange oversættere i dag arbejder direkte i ordregivernes systemer og ikke individuelt med egne programmer og databaser.

Det foreliggende ordbogsprojekt er dog et af dem, der har kørt videre frem til i dag med successive opdateringer og nyudvikling af online-ordlister til fri afbenyttelse. Der har været en løbende indsamling af termer i det daglige arbejde og en målrettet tematisk termindsamling i forbindelse med større opgaver. Somme tider er udviklingsarbejdet også sket i samarbejde med kolleger, ordregivere og virksomheder og organisationer. Enkelte meget interesserede privatpersoner har også bidraget med materiale. Desuden er der gennem årene blevet opbygget et stort referencebibliotek med fagbøger.

I 2008 modtog projektet økonomisk bidrag fra Nordplus til at videreføre og udvikle projektet, og projektet er blevet refereret af flere, bl.a. af Pálfi (2011) og Kristensen (2012), og har ført til flere artikler (Hansen 2016, 2019 og 2021).

Imidlertid er søgningen på vores tematiske termlister mellem dansk og svensk faldet betydeligt i de sidste år. Der er kommet flere gratis online ordbogsresurser mellem dansk og svensk, f.eks. Svensk-Dansk Ordbog og

sprogbro.org. Derfor blev der i 2020 indgået en licensaftale med det svenske firma WordFinder, hvilket førte frem til en publicering af vores tematiske ordlister i et brugervenligt og samlet format.

8. Ordbogens strukturer

Ordbogens oplæg er udviklet i samarbejde med WordFinder med udgangspunkt i et tidligt udkast til et mobilapp-projekt, foreslået af EMP AB – Erlandsen Media Publishing. Inspiration til den tematiske opbygning er også hentet fra Dahlerup (1919) og Bendz (1965). Strukturen vises i Tabel 3.

TABEL 3. Oversigt over de vigtigste strukturer.

Søgefeltet	I søgefeltet angives den term, der skal oversættes. Det er også muligt at angive ordkombinationer i søgefeltet (eller vælge dem i den smalle scroll-liste til venstre i displayet, se Figur 1), f.eks. ”blindplugg för kontakthölje” og ”straffet lindrades till böter”. Fritekstsøgning er mulig.
Fagordsinddeling	Ordbogen inddeler fagordene i to kategorier: en overordnet kategori (f.eks. biologi) og en underordnet kategori (f.eks. svampe).
Forklarende tekst	Forklarende tekst i parenteser er brugt i visse tilfælde, f.eks. ”bolde [legetøj]”.
Latinske navne	Latinske videnskabelige navne er angivet for ca. 30.000 zoologiske og botaniske arter. Løser problemer med inkonsekvente danske og svenske navne.
Skift af sprogretning	Ordbogen foreligger i en svensk-dansk og en dansk-svensk version. Ønsket version aktiveres før søgning. Langt de fleste opslagsord kan søges i begge sprogretninger.
Advarsler om lumske ord	Der er indsat ca. 2.000 generelle advarsler om lumske ord. De angives i søgeresultatet med en standardformulering: ”NB. Kan være lumsk, betyder ikke altid helt det samme som følgende danske ord (eller er helt forskellig):”. Efter kolon angives det pågældende ord.
Tilvalg af andre terminologiske resurser	WordFinders andre fagordbøger (f.eks. IATE og digitaliserede TNC-termlister) kan aktiveres og automatisk indgå i søgningen.

9. Ordbogens temaer

Ordbogen har 16 overordnede og 60 underordnede temaer i sin nuværende udformning (se Tabel 4).

TABEL 4. Oversigt over ordbogens temaer.

Overordnet niveau
Biologi, EU-termer, fuglenavne, hverdagsord, kemi, landbrug, limstoffer, lystfiskeri, medicin, musik, spildevand, talemåder, teknik, tekniske betegnelser, produkter/ydelser og økonomi.
Underordnet niveau
Alment sprog **–** basisordforrådet i dansk og svensk (ca. 25.000 opslagsord) samt underordnede fagområder, herunder ejendomshandel, socialt arbejde, skoletermer, mødetermer, geografi, arbejdsmarkedsord, fødevarer, beklædning, køkkenudstyr, kontor, lumske ord, modsat køn, chatsprog, forkortelser, farlige ord, ordpar og slang.
Teknik **–** data, bilteknik, byggeteknik, værktøj, arbejdskøretøjer, plasmaskæring, svejseteknik, skruer og bolte, Microsoft-ord, kemikalier, busteknik, maskiner, metallurgi, sejlbådsteknik, generel maskinteknik, vvs, affaldsteknik, bygge & anlæg, el, jernbane, tilsætningsstoffer, knusemaskiner, mekanikerord, sikkerhed, E-numre, S-sætninger, R-sætninger og grundbegreber.
Biologi og medicin **–** kulturplanter, skadedyr, sygdomme, ukrudt, botaniske termer, typiske plantearter (for nåleskov, løvskov, eng, kulturlandskab, hede, mose, kær, sø, vandløb, strand, fjeld), svampe, fugle, danske fuglesynonymer, vandlevende dyr, skade- og nytteinsekter, lægemidler, forkortelser, diagnostik og kirurgi.

10. To praktiske eksempler

Figur 1 viser søgeresultatet i en oversættelsessituation vedrørende den svenske term *skyddsräcke*. Termen forekommer i teksten "Plattform och trappa utan skyddsräcken". Problemet er at finde en dansk term, som passer ind her. På svensk kan man bruge termen *skyddsräcke* i forbindelse med både platform og trappe, men på dansk vil man nok vælge forskellige termer til trappe og platform. I søgeresultatet ses tre forskellige oversættelsesforslag under kategorien "sikkerhed", nemlig *lønning*, *rækværk* og *håndliste*. Til

platformen kan man bruge termen *rækværk,* og til trappen kan man evt. overveje termen *håndliste* (men måske er det endnu bedre med *gelænder*). Det er ikke angivet som et alternativ i resultatet i Figur 1, men man har jo lov til at tænke selv. Alle søgemuligheder er dog ikke blevet udnyttet her. Hvis man før søgningen havde valgt at aktivere søgning i WordFinders andre resurser (f.eks. IATE og TNC-ordbøgerne), ville alternativet *gelænder* også have været vist i søgeresultatet. Et oversættelsesforslag kunne således være: "Platform uden rækværk og trappe uden gelænder".

FIGUR 1. Resultatet af en søgning på termen *skyddsräcke* i en oversættelsessituation.

Figur 2 viser en søgning i en efterredigeringssituation. Man ønsker at vurdere, om termen *dam* er korrekt oversat i den tekst, man efterredigerer, f.eks. "dammen bestod mest af sand", som er grammatisk korrekt, men ikke føles helt logisk. Først bruges ordbogen i dansk-svensk sprogretning. I søgeresultatet advares der om, at *dam* er et lumsk ord, og at der er risiko for fejloversættelse med hensyn til svensk *damm* og *dam*, men i opslaget er der også anden information, som kan bruges til at vurdere, om der kan være tale om en oversættelsesfejl i dette tilfælde. Næste skridt er at aktivere ordbogen i svensk-dansk sprogretning og søge på svensk *damm* og *dam* for at klarlægge de danske oversættelsesalternativer. Herved fremkommer, at der er et oversættelsesalternativ, hvis "dammen" faktisk er en fejloversættelse: "dæmningen bestod mest af sand". Sammenhængen må derefter afgøre, hvad der er mest korrekt.

FIGUR 2. Resultatet af en søgning på termen *dam* i en efterredigeringssituation.

11. Afslutning

Der er fortsat et stort behov for fagsproglig oversættelse mellem dansk og svensk, og der er fortsat behov for udvikling af gode professionelle ordbogsresurser. Der er dog sket en stor forandring i den måde, arbejdet udføres på. Vi lever i en tid, hvor det praktiske oversættelsesarbejde ofte handler om efterredigering af tekster, der er automatisk oversat ved hjælp af oversættelsesværktøjer. Som det er påpeget af Leroyer & Simonsen (2019), bør oversættelsesordbøger derfor ændres til i højere grad at understøtte efterredigering med situationstilpassede datakategorier. Her kan hårdt strukturerede flerfagsordbøger med høj grad af oversættelsesækvivalens blive særligt værdifulde.

Den flerfagsordbog, som nu er blevet lanceret, er et forsøg på at skabe en opdateret, aktualiseret fagleksikografisk resurse ved at tage udgangspunkt i de fagområder, hvor der faktisk lige nu er et marked for faglig oversættelse mellem dansk og svensk. Den grundlæggende idé bag ordbogen er, at samspillet mellem brugerens oversættelsesmæssige kompetencer og vidensniveau på den ene side og en præcis og kortfattet information i søgeresultaterne på den anden side skal munde ud i korrekte oversættelser.

Der er foretaget flere tilpasninger for at imødekomme en fagoversætters generelle behov og ønsker i forbindelse med oversættelse og efterredigeringsarbejde. For det første er der efterstræbt en tydelig og detaljeret opdeling af indholdet med temaer i to niveauer, som gør det muligt at få et hurtigt overblik, hvis en term har forskellige oversættelsesvariationer. For det andet er informationsmængden i søgeresultatet reduceret til et minimum, idet det forudsættes, at brugeren ikke er en lørner. For det tredje er der lagt vægt på at identificere, hvilke termer og ordkombinationer der kan være lumske i fagsprog ved oversættelse mellem dansk og svensk, hvilket få (eller måske ingen) har prøvet at sammenstille tidligere.

Med hensyn til det sidstnævnte er der et konkret leksikografisk udviklingspotentiale. Der vil være gode chancer for, at en målrettet indsamling af lumske fagord (f.eks. *beskyttet bolig*, *børnebidrag*, *stensætning*, *snedkerhammer* og *skovsneppe*) og implementering af dem i efterredigeringssoftware kan føre til en automatiseret og effektiviseret lokalisering af fejloversættelser i fagsprog.

Referencer

Ordbøger

Dahlerup, Verner 1919. *Svensk-dansk ordsamling*. København og Kristiania. Gyldendalske Boghandel. Nordisk Forlag.
Dansk-Svensk Ordbok. <https://www.wordfinder.com/>. Besøgt marts 2022.
IATE (Interactive Terminology for Europe). <https://cdt.europa.eu/en/iate>. Besøgt marts 2022.
Linder, Bernhard 1965. *Svensk dansk teknisk ordbog*. København: Munkgaards Forlag.
Orlando, Galindo 2006. *Svensk-dansk fackordbok*. København: GTO.
Proz.com. <https://www.proz.com/>. Besøgt marts 2022.
Rikstermbanken. <http://www.rikstermbanken.se/>. Besøgt marts 2022.
sprogbro.org. <https://sprogbro.org/>. Besøgt marts 2022.
Svensk-Dansk Ordbog. <https://ordnet.dk/sdo/>. Besøgt marts 2022.
Wordfinder. <https://www.wordfinder.com/>. Besøgt marts 2022

Anden litteratur

Bendz, Gerhard 1965. *Ordpar*. Stockholm. Norstedt & Söners Förlag.
Bergenholtz, Henning & Sven Tarp (red.) 1994. *Manual i fagleksikografi. Udarbejdelse af fagordbøger. Problemer og løsningsforslag*. Herning: Systime.
Fuertes-Olivera, Pedro A. & Sven Tarp 2014. *Theory and practice of specialised online dictionaries: Lexicography versus terminography*. Berlin/New York: De Gruyter.
Hansen, Poul 2016. Hvordan har netbrugernes præference for forskellige emnespecifikke ordbøger ændret sig i perioden 2006-2014? I: Asgerd Gudiksen & Henrik Hovmark (red.), *Nordiske studier i leksikografi* 13. København: Nordisk forening for leksikografi, 131-141.
Hansen, Poul 2019. Statistisk modellering og prognosticering af efterspørgslen efter netordlister. *LexicoNordica* 26, 55-73.
Hansen, Poul 2021. Hvordan har brugen af mobiltelefoner indvirket på efterspørgslen efter netordlister? I: Caroline Sandström, Ulla-Maija Forsberg, Charlotta af Hällström-Reijonen, Maria Lehtonen & Klaas

Ruppel (red.), *Nordiska studier i lexikografi* 15. Helsingfors: Nordisk förening för lexikografi, 125-133.

Leroyer, Patrick 2018. Bruger- og ekspertinddragelse ved udarbejdelse af online (fag)ordbøger: det kooperative princip i leksikografien. I: Gudiksen, Asgerd & Henrik Hovmark (red.), *Nordiske studier i leksikografi 13*. København: Nordisk forening for leksikografi, 177-190.

Kristensen, Kjeld. 2012. Tre svensk-danske ordbøger på nettet. *LexicoNordica* 19, 273-293.

Leroyer, Patrick & Asta Høy 2016. Vinsmagningsordbogen Œnolex Bourgogne. En milepæl i pragmatisk fagleksikografi. I: Ruth Vatvedt Fjeld & Marit Hovdenak (red.), *Nordiske studier i leksikografi* 12. Oslo: Nordisk forening for leksikografi, 287-302.

Leroyer, Patrick & Liselotte Kruse 2012. Ejendomsordbogen fransk/dansk: ny integreret e-ordbog. I: Birgit Eaker, Lennart Larsson & Anki Mattisson (red.), *Nordiska studier i lexikografi* 11. Lund: Nordisk förening för lexikografi, 405-417.

Leroyer, Patrick & Henrik Køhler Simonsen 2019. Google Translate: trussel eller redning for oversættelsesbøger? *LexicoNordica* 26, 95-115.

Pálfi, Lorand-Levente 2011. *Leksikon over ordbøger og leksika*. København. Frydenlund.

Digitalisering af Jysk Ordbogs seddelsamling

Inger Schoonderbeek Hansen, Mette-Marie Møller Svendsen & Kristoffer Friis Bøegh

The historical, lexicographical documentation project, the Dictionary of the Jutlandic Dialects, *Jysk Ordbog*, is being edited on the basis of an extensive, unique, and *physical* collection. This collection comprises over three million paper slips containing excerpts of written and spoken linguistic data, reflecting the traditional culture and social conditions of the rural population of Jutland, Denmark, around 1850–1920. This article presents and discusses a pilot project in which almost 56,000 individual paper slips were scanned and digitized. The experiences of the project provide insights into how *Jysk Ordbog* and its different types of end-users will be able to benefit from digitization, and, in a more general documentary, lexicographical perspective, how digitization supports the project's overall goal of preserving cultural heritage.

NØGLEORD: dialektleksikografi, digitalisering, seddelsamling, Jysk Ordbog, kulturarv

1. Indledning

Den historiske dokumentationsordbog *Jysk Ordbog* (herefter JO) redigeres på baggrund af et omfattende, unikt og *fysisk* kildemateriale. Det består af ca. 3,1 millioner excerpter af forskellige typer skrift- og talesproglige kilder der giver indblik i den jyske landbefolknings kultur- og samfundsforhold i perioden ca. 1850–1920. JO's kildemateriale er imidlertid ikke digitalt og matcher derfor ikke nutidens leksikografi (og sprogvidenskab generelt) hvor der i udgangspunkt tages afsæt i *digitale* tekster og korpora (jf. fx Svensén 2004, Atkins & Rundell 2008, van Keymeulen 2018). Overordnet set kan nutidens syn på tekst opsummeres således:

> [T]exts in the 21st century are digitized. We digitize cultural heritage and literary classics, and new texts within mass communication, education, social media, journalism and literature are by default produced in digital formats (Bertelsen & Tannert 2021:i).

I nærværende artikel præsenteres arbejdet med et pilotprojekt hvor ca. 56.000 sedler fra JO's seddelsamling blev scannet og bearbejdet digitalt.[1] Overordnet set kommer artiklen ind på hvad digitalisering af seddelsamlingen betyder, og hvem der kan have gavn af den. De opnåede erfaringer med pilotprojektet giver et indblik i hvordan JO og dens brugere fremadrettet vil kunne drage nytte af digitaliseringen, og dermed hvordan denne understøtter projektets mål med bevaring af kulturarv og indgang til lingvistiske studier. Mens Hansen (2020) gav en statusrapport for JO med afsæt i spørgsmål om *hvem* ordbogen redigeres af og til, samt *hvorfor* det traditionelle videnskabelige ordbogsarbejde fortsat har og skaber værdi, supplerer nærværende artikel med afsæt i digitaliseringsprojektet med *hvordan* brugerne vil kunne tilgå det omfattende materiale.

I afsnit 2 skitseres JO's heterogene kildemateriale, og i afsnit 3 redegøres for pilotprojektets resultater. Det gælder såvel arbejdet med at ekstrahere metadata fra det scannede materiale, som hvilke konkrete udfordringer scanning af de redigerede og uredigerede sedler gav. I afsnit 4 drøftes digitaliseringens formål, dvs. hvilke nye forskningsperspektiver og andre anvendelsesmuligheder en digitaliseret seddelsamling vil kunne medføre. Desuden gives nogle eksempler på både simple og avancerede søgninger, der vil kunne fortetages af diverse brugertyper, og der vil kunne give større og mere fleksible muligheder for at undersøge opslagsord af specifik sproghistorisk og/eller kulturhistorisk værdi. I afsnit 5 kommer vi med nogle afsluttende bemærkninger samt perspektiverer til fremtidige tiltag.

2. Jysk Ordbogs kildemateriale

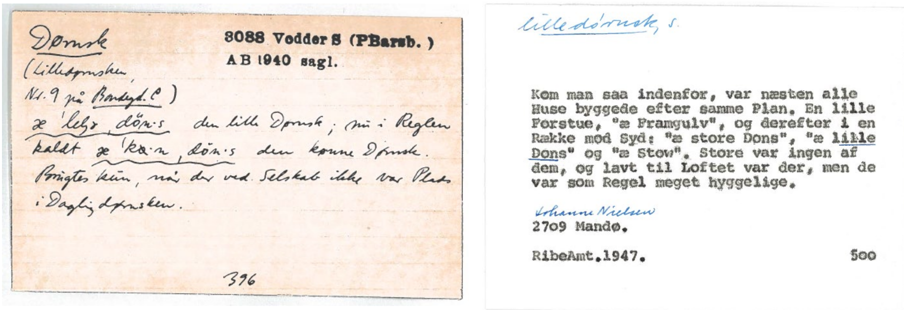
Det omfattende, empiriske indsamlingsarbejde til JO påbegyndes i 1932 af professor i dansk sprog ved Aarhus Universitet Peter Skautrup (1896–1982). JO er derfor et empirisk-videnskabeligt funderet projekt der afspejler dansk kulturarv sådan som den sprogligt er indlejret hos den jyske landbefolkning i perioden ca. 1850–1920, dvs. indtil afdialektaliseringen i Danmark tog fart som følge af bl.a. landbrugets mekanisering samt samfundets industrialisering og urbanisering (jf. fx Bøegh et al. 2023). De første prøvehæfter af ordbogen udkom 1970–1979, og siden 2000 er JO udkommet

1 Stor tak til Kristoffer Nielbo og Peter Bjerregaard Vahlstrup, Aarhus Universitet samt studentermedhjælper Agnes Boel Nielsen for deres bidrag til pilotprojektet.

som online ordbog på www.jyskordbog.dk. Siden 2017 anvendes redaktionsværktøjet iLEX, og i skrivende stund (marts 2023) er 53 % af JO færdigredigeret og publiceret på https://jysk.au.dk/jyskordbog. Selvom JO publiceres online, findes selve kildematerialet fortsat i fysisk form (jf. Arboe 2003, Sørensen 2019, Hansen 2020, jysk.au.dk/jyskordbog).

JO's kildemateriale kan inddeles i tre hovedgrupper:

1) JO's seddelsamling eller centralkartotek (jf. JO, Ordbogens kilder) indeholder ca. 3,1 millioner papirsedler med excerpter fra bl.a. skønlitteratur på jysk dialekt, jysk typografi og lokalhistorie, dialektoptegnelser ved filologer mv. De enkelte papirsedler indeholder ord, udtalegengivelser, ordforbindelser og/eller korte citater af autentisk jysk dialekt. Sedlerne udgør et i høj grad heterogent materiale; nogle sedler er skrevet i hånden, andre på skrivemaskine, mens andre igen er indtastet på computer og derefter udprintet. De afspejler dermed mange årtiers dialektindsamling, forskellige faglige traditioner, varierende grader af systematik og akribi etc.; sedlernes indhold er ligeledes særdeles heterogent (se Figur 1).


8088 Vodder S (PBarsb.)
AB 1940 sagl.
396

lilledørnsk, s.
Kom man saa indenfor, var næsten alle Huse byggede efter samme Plan. En lille Forstue, "æ Framgulv", og derefter i en Række mod Syd: "æ store Dons", "æ lille Dons" og "æ Stow". Store var ingen af dem, og lavt til Loftet var der, men de var som Regel meget hyggelige.
Johanne Nielsen
2709 Mandø.
RibeAmt.1947. 500

FIGUR 1. To eksempler fra seddelsamlingen til opslagsordet *lille·dørns*: dialektologen Anders Bjerrums seddel fra Vodder (til venstre), og meddeler Johanne Nielsens beretning "Livet paa Mandø" i Fra Ribe Amt, 1947 (til højre).

Sedlernes heterogenitet taget i betragtning må man regne med "mange typer sproglige data på en ordbogsseddel", som Henrik Hovmark (2011:301) fra *Ømålsordbogen*[2] formulerer det, hvilket fremhæver vigtigheden af at have et kritisk blik på sedlernes repræsentativitet. Sedlerne gengiver derfor "et udsnit af virkeligheden som skal vurderes kritisk som kildegrund-

2 Jf. Gudiksen (2021) for en nylig oversigt over *Ømålsordbogens* arbejde.

lag" (2011:305). De kan således aldrig anses som 100 % repræsentative, men de giver et kvalificeret bud på hvordan landbefolkningen talte.

2) JO's spørgelistesamling består af spørgelister udsendt 1949–2014, besvaret af mellem 300 og 1.000 meddelere født i perioden ca. 1880–1920. Denne kildetype bidrager primært til beskrivelsen af den geografiske udbredelse af ords betydning og/eller udtale, i JO illustreret ved hjælp af udbredelseskort. Spørgelistesamlingen er ligeledes excerperet og skrevet på sedler der indgår i seddelsamlingen. Endelig er spørgelistesamlingen delvist digitaliseret og kan tilgås via en lokal database på Peter Skautrup Centret (herefter PSC) i Århus.

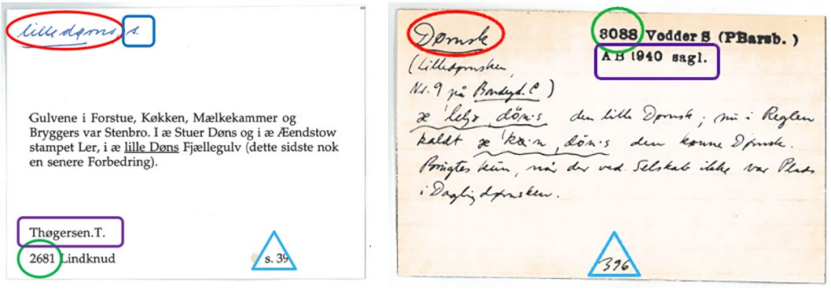
3) JO's båndsamling består primært af optagelser fra hele landet indsamlet i perioden 1971–1976 og med dialekttalende født omkring 1900. En stor del af den blev indsamlet på basis af flere samarbejdsprojekter med Center for Dialektforskning, Københavns Universitet. Båndsamlingen er ligesom spørgelistesamlingen digitaliseret; en digital kopi af den opbevares af Det Kongelige Bibliotek, Aarhus (jf. Goldshtein & Puggaard 2019). Kun en mindre del af båndsamlingen er excerperet, dvs. transskriberet og udskrevet på sedler der indgår i JO's seddelsamling.

3. Pilotprojektet

Pilotprojektet med digitaliseringen af JO's seddelsamling er et samarbejde mellem PSC og Center for Humanities Computing Aarhus (herefter CHCAA). For PSC tjener digitalisering af den omfattende fysiske, og sårbare, seddelsamling et bevaringsformål, mens en digitaliseret seddelsamling i højere grad vil kunne stilles til rådighed for både fagfolk og alment interesserede end den fysiske samling kan i dag. Den vil yderligere kunne give bedre online søgeværktøjer for ordbogsredaktørerne, bl.a. vil man kunne lave krydssøgninger på metadata, fx ordklasse og kilder. CHCAA ville med pilotprojektet gerne teste kapacitet og brugbarhed af en specialscanner, samt tilpasse software de har udviklet til scanningsarbejdet. Dette ville give CHCAA gode erfaringer i forhold til fremtidige scanningsprojekter som vil kunne tilbydes til andre fysiske samlinger i bog- eller seddelform.

3.1. Metadata

JO's seddelmateriale er som nævnt heterogent. På en eksemplarisk seddel indgår et opslagsord, ordklasse, et topografisk nummer, navn på kilden, et excerpt og evt. et sidetal hvis excerptet er fra en trykt kilde. Alle sedler besidder dog langt fra altid alle nævnte metadata, og placeringen af de enkelte metadata på sedlen varierer meget (se Figur 2). Desuden er mange af sedlerne håndskrevne. Dette giver nogle udfordringer i forhold til en automatiseret afkodning af data fra sedlerne. Et af digitaliseringens mål er at *computer vision* anvendes til at aflæse udvalgte metadata fra seddelsamlingens enkelte sedler. I de fleste tilfælde vil computeren kunne tilgå JO's eksisterende lukkede inventarlister, fx redaktionens lister over kilder, topografiske numre mv.



FIGUR 2. To repræsentative eksempler fra JO's seddelsamling til opslagsordet *lille·dørns*. Den røde cirkel viser metadata opslagsord, den blå firkant ordklasse, den grønne cirkel topografisk nummer, den lilla firkant kilde, og den lyseblå trekant kildens sidetal.

3.2. Scanningsresultater

Til pilotprojektet benyttede vi en seddelscanner med en kapacitet på 250 sedler i minuttet. Dertil brugte vi software specialfremstillet af CHCAA til at sortere det scannede materiale og indtaste udvalgte metadata. En studentermedhjælper stod for håndteringen af seddelmaterialet og scanningen, fx skulle hun indtaste udvalgte metadata. Undervejs i forløbet førte hun logbog over hvilke skuffer og typer materiale hun scannede, hvor lang tid det tog, og hvilke problemer hun stødte på i løbet af processen. Til pilotprojektet stræbte vi efter at vælge et så forskelligartet materiale

som muligt da vi ville undersøge hvilke faktorer der kunne påvirke scanningsprocessen. Fx valgte vi nogle skuffer med få sedler per opslagsord, dvs. mange, men korte opslagsord som studentermedhjælperen skulle behandle, og sammenlignede disse erfaringer med skuffer med blot et enkelt eller få opslagsord, dvs. med flere tusinde sedler per opslagsord. Dertil lavede vi også stikprøver i forhold til redigeret og uredigeret seddelmateriale. Det heterogene seddelmateriale gav umiddelbart en del udfordringer. Her følger nogle eksempler:

1) Seddelmateriale med trykt information på begge sider. Scanneren er indrettet til at kunne registrere tekst på begge sider af en seddel, men da papirkvaliteten på flere af sedlerne er helt tynd, ville det resultere i en del fejlscanninger og i værste fald i den dobbelte arbejdsmængde hvis samtlige sedler skulle scannes to gange. Løsningen blev at scanneren blev indstillet til kun at scanne én side hvorfor studentermedhjælperen skulle monitorere scanningsprocessen og holde øje med sedler med tryk på begge sider.

2) Seddelmateriale sammensat af to (eller flere) stykker papir. Scanning af sedler bestående af flere sammenklipsede eller -hæftede stykker papir kan resultere i papirstop, eller i værste fald iturevne sedler. En løsning på problemet var at studentermedhjælperen overførte data fra sedlen til ét stykke papir, fx ved at scanne sedlerne på en almindelig scanner og efterfølgende klippe det til.

3) Uredigeret seddelmateriale. Til trods for at seddelsamlingen er struktureret både alfabetisk og topografisk, kan uredigeret materiale mangle finsortering og derfor stå i forskellige skuffer; det gælder fx materiale til ordformerne *lodden* og *lådden* der begge hører til opslagsordet *lodden*. Visse homonymer er heller ikke opdelt endnu hvorfor studentermedhjælperen skulle foretage flere beslutninger i scanningsprocessen.

JO's 3,1 millioner sedler er fordelt på 1.656 skuffer, dvs. ca. 1.900 sedler per skuffe, der varierer i størrelse. Studentermedhjælperen brugte 74 timer på projektet hvori hun scannede 25,5 skuffer med 55.897 sedler, fordelt på 2.557 opslagsord. Det gav os altså ca. 755 indscannede sedler i timen eller 1/3 skuffe i timen. Det betyder at vi i løbet af pilotprojektet fik scan-

net ca. 2 % af seddelsamlingen. En scanning af hele seddelsamlingen vil derfor kræve ca. 3.700 studentertimer. Selvom timeantallet er højt, og vi sandsynligvis vil møde nye problemstillinger ved en opskalering af projektet, så har vi løst mange udfordringer: Redaktionen har i fællesskab diskuteret og fastlagt tilgange til at håndtere seddelmateriale der giver ekstra udfordringer, og som derfor ikke kommer til at kræve en masse ressourcer fremover. CHCAA har justeret softwaren efterhånden som studentermedhjælpen er stødt på noget der yderligere kunne optimere indtastningsarbejdet. Arbejdsgangen er blevet optimeret undervejs i projektet, og gode vejledninger til processen er blevet udfærdiget, så arbejdet ved opskalering hurtigt kan igangsættes. Ved en opskalering vil der også blive arbejdet med automatisk aflæsning og fejltjek af metadata (jf. afsnit 3.1).

4. Hvilket formål har en digitaliseret seddelsamling?

De enkelte opslag i JO fungerer, som Asgerd Gudiksen påpeger i sin artikel om *Ømålsordbogens* samlinger som sproghistorisk kilde, som en "indgang til [den bagvedliggende] samling" (Gudiksen 2021:25). Hver artikel baserer sig derfor på forskning der formidles både til den alment interesserede bruger og til fagfællen, lokalhistorikeren etc. (jf. Svendsen 2023). I nærværende afsnit fremhæves digitaliseringens potentiale i et forsknings- og brugerorienteret perspektiv (jf. desuden Hovmark 2023).

4.1. Fremtidig brugerflade

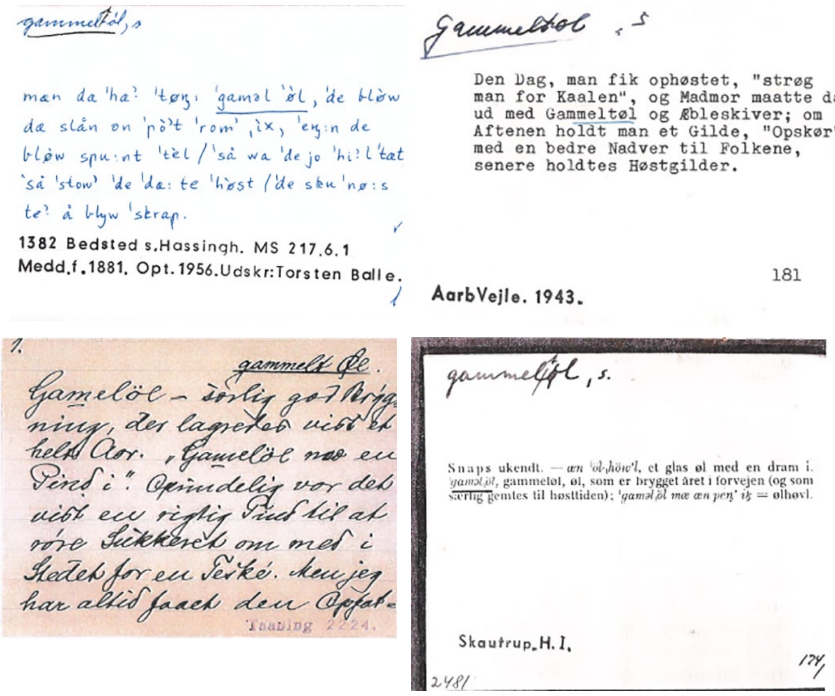
En digital seddelsamling vil på den ene side gavne JO's redaktionsarbejde, især hvis den kan udformes som et dynamisk værktøj hvor uredigerede sedler kan sorteres og opdateres (jf. Hovmark & Gudiksen 2018). På den anden side vil den også kunne bruges af kolleger, ikke mindst som grundlag for egne sproghistoriske og kulturhistoriske studier. Endelig vil den almindeligt interesserede bruger kunne tilgodeses med oplysninger om dansk talesprogs historie, struktur og kulturafspejling.

JO's brugere bør kunne få adgang til en brugerflade hvor de kan søge på forskellige typer data både som simple eller mere avancerede søgninger. Søgninger skal kunne trække på oplysninger i de to tilgængelige databaser: redigeringsplatformen iLEX og den fremtidige digitaliserede seddelsamling. Løsningen skal give mulighed for at vise seddelmaterialet, lige-

som brugerne på sigt via ordbogens kortmateriale skal kunne se et ords udbredelse, sådan som de er vant til i JO. Dermed vil det materiale redaktionen igennem årtier har indsamlet og bearbejdet leksikografisk, kunne nå ud til såvel en bredere offentlighed som det mere snævre lingvistiske fagmiljø. Nedenfor reflekteres over disse to brugertyper med afsæt i nogle eksempler.

4.2. Simple søgninger

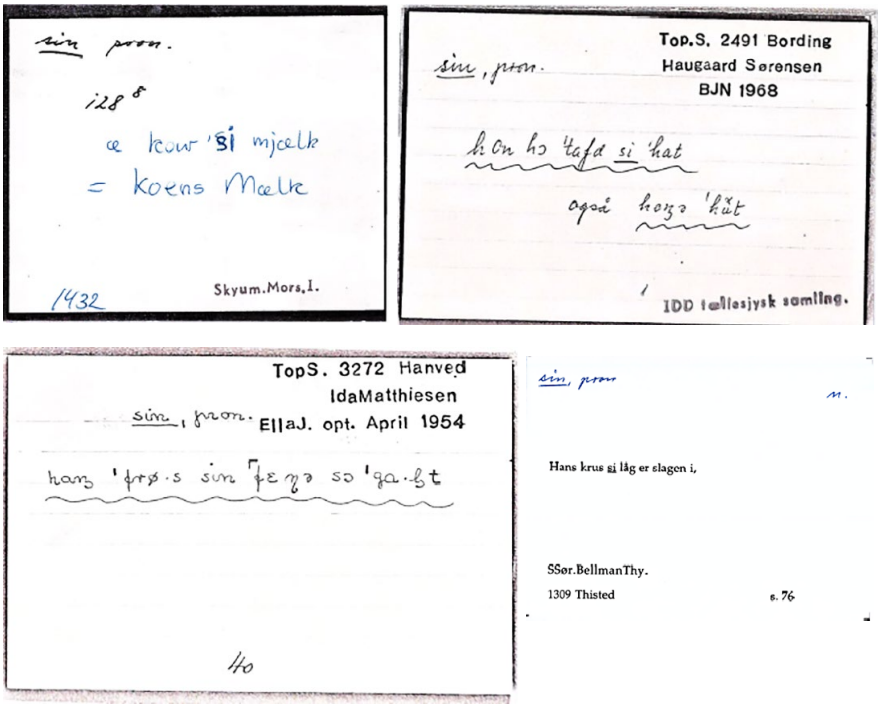
Hvis en alment interesseret bruger ønsker at vide mere om betydningen og udbredelsen af et specifikt ord, kan vedkommende slå ordet op i JO. Ifølge JO betyder fx ordet *gammeløl* 'lagret, stærkt øl', og der oplyses yderligere: "som regel brygget i marts (el. tidligere) og anvendt i høst- og juletiden, stedvis også ved gilder; ofte tilsat sukker (el. kandis) og brændevin (el. rom)". Betydningsangivelsen følges op af fem citater, dvs. belæg som redaktøren har udvalgt fra seddelsamlingen (jf. også Gudiksen 2021:25). Hvis brugeren derudover kunne få adgang til en digitaliseret version af de relevante sedler, inkl. de øvrige eksempler på *gammeløl* som redaktionen af bl.a. formidlingsmæssige grunde har fravalgt fra selve ordbogsartiklen, ville vedkommende have mulighed for selv at gennemgå opslagsordets belæg. Brugeren kunne ønske at vide mere om hvordan man bryggede øl i det traditionelle jyske landbosamfund. Da artiklen *øl* endnu ikke er redigeret i JO (som er nået til *l-*), ville en søgning i JO's digitaliserede seddelsamling på opslagsordet *øl*, evt. inkl. ordklasseangivelse, give et relativt stort antal excerpter, fx:



FIGUR 3. Et udvalg af excerpter under et forestillet søgeresultat af opslagsordet *gammel·øl*.

Sedlerne ville være topografisk ordnede, men som det fremgår ovenfor, kan excerpterne af forskellige årsager være svære at læse (jf. afsnit 2). Derfor bør sedlerne suppleres med oplysninger, linkmuligheder mv., så brugeren vil kunne få et indblik i fx kildernes alder, proveniens og pålidelighed.

Man kan ligeledes forestille sig en situation hvor en fagfælle, fx en lingvistkollega, gerne vil vide mere om et jysk grammatisk fænomen, fx brugen af de refleksive pronominer *hans*, *hendes* og *sin*. Lingvistens søgning i JO resulterer i oplysninger under hhv. opslagsordene *han* og *hun* hvor oplysninger om det refleksive *hans* hhv. *hendes* er bragt under artiklens allersidste betydningsangivelse. Her viser eksempler at *hans* og *hendes* i jysk kan referere til sætningens subjekt når det er hhv. *han* og *hun*. Dette mønster afviger fra rigsdansk hvor det refleksive pronomen *sin* bruges, hvilket nævnes i ordbogsartiklens afsnit om etymologi. Hvis lingvisten gerne vil vide mere om brugen af det endnu uredigerede *sin*, fx om det overhovedet bruges i jysk, så vil hun kunne søge på opslagsordet *sin* i den digitale seddelsamling og finde et relativt stort antal excerpter fra samlingen, fx:



FIGUR 4. Forestillet søgeresultat for pronomenet *sin* – et udvalg af excerpter.[3]

Også dette søgeresultat bør suppleres med oplysninger der gør det muligt for lingvisten at forholde sig lige så kildekritisk til materialet som JO's redaktører.

4.3. Avancerede søgninger

Brugerne, uanset om de er ikke-lingvister eller lingvister, skal også kunne foretage avancerede søgninger i den digitaliserede seddelsamling. Til dette formål er det afgørende at kunne afgrænse søgninger ved hjælp af filtre. Umiddelbart vurderer vi følgende filtre som mest relevante: opslagsord, ordklasse, topografisk nummer og kilde. Senere kan der evt. tilføjes yder-

3 Bent Jul Nielsen lavede en større undersøgelse af pronomenet *sin* i jysk (Nielsen 1986). Resultaterne viste at *sin* kan optræde som refleksivt pronomen i jysk, men at det da som regel henviser til subjekter der ikke er *han* eller *hun*, jf. fx *e kow' stå'r i **si** bo·'s* (optegnelse af Torsten Balle i Thorsted, Thy, ca. 1960) eller *den (dvs. vædderen) havde jo gjort **sin** tjeneste* – hvorfor den skulle slagtes (Anders Bjerrum, optegnelse fra Vodder, Vestsønderjylland, ca. 1930'erne).

ligere filtre som fx emne og semantisk kategori. Mens simple søgninger på metadata som opslagsord plus ordklasse allerede foreligger i pilotprojektet (jf. afsnit 3), kunne den almindelige bruger anvende et filter som topografisk nummer, dvs. en søgning på geografisk variation inden for det jyske område, for at skærpe sit spørgsmål til fx "Brugte man betegnelsen *gammeløl* i Vendsyssel?" Brugerne finder oplysninger om hvad de specifikke topografiske numre refererer til i en lukket liste; listen indgår i en af JO's redaktionsmanualer og skal blot gøres tilgængelig. Lingvisten kunne bruge filteret 'topografisk nummer' til fx at undersøge om der findes forskel på brugen af *hans* og *sin* i hhv. vestjysk og østjysk, måske ud fra en hypotese om at østjysk i flere tilfælde strukturelt stemmer overens med rigsdansk, jf. fx artikelbrug: østjysk *hus-et* og vestjysk *æ hus*. Dette filter bør på sigt kobles sammen med det omfattende kortmateriale der er tilgængeligt i JO. Filteret kilde kan bl.a. muliggøre søgning på historisk variation i materialet. Hvis brugeren har fundet et ord i en specifik kilde, fx hos en dialektforfatter som Jeppe Aakjær, skal vedkommende have mulighed for udelukkende at kunne søge i den del af JO's seddelsamling der består af excerpter fra Jeppe Aakjærs tekster. Filtret skal hjælpe lingvisten i sin søgning til fx at kunne datere et specifikt sprogligt fænomen. En liste over JO's kilder fremgår allerede på JO's hjemmeside, og brugeren kan finde oplysninger om kildens alder, hvad den beskriver etc.

Fremtidige søgemuligheder kunne fx også inkludere søgning efter emne, som pt. hverken eksisterer som tags for seddelsamlingen eller i JO som sådan, men som anvendes i visse andre ordbøger. Ydermere kunne man forestille sig søgninger efter bestemte semantiske kategorier; dette er pt. heller ikke muligt, men kunne opnås ved at gøre det muligt udelukkende at søge i opslagsordets betydningsafsnit. Dette ville umiddelbart være muligt i det færdigredigerede men ikke i det uredigerede seddelmateriale.

5. Afslutning

Digital humaniora og digitaliseret kulturarv er kommet for at blive, og de bidrager til nye, forskningsmæssige indsigter. Digitalisering af arkiv- og seddelsamlinger har gennem de seneste 25-30 år fundet sted på forskningsinstitutioner rundtom i Europa. I en nordisk kontekst var et af de nok tidligste projekter det norske Dokumentationsprojekt (Ore & Kristiansen 1998), mens et af de senere projekter er digitalisering af *Ømålsordbogens*

seddelsamling, der dog endnu mangler en digital infrastruktur i form af en brugerflade (jf. Hovmark & Gudiksen 2018). Formidlingsmæssigt vinder vores projekt ved at unikt og empirisk-videnskabeligt materiale om det gamle landbosamfund i Jylland kan nå ud til et langt større publikum med såvel faglig som almen interesse. Det både form- og indholdsmæssigt heterogene materiale fordrer dog vejledninger til de enkelte kilders optegnelser, korte emnebeskrivelser, så brugerne bedre kan forstå og bruge materialet. Pilotprojektet har givet en smagsprøve på hvad vi kan forvente ved en opskalering af pilotprojektet med digitalisering af JO's seddelsamling. En fremtidig målsætning er ydermere at koble de eksisterende og sammenlignelige ordbogsressourcer i såvel Danmark som i Norden (jf. også Svendsen 2023).

Litteratur

Arboe, Torben 2003. Jysk Ordbog på internettet. *Nordiske Studier i Leksikografi* 6, 31–41.

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

Bertelsen, Ulf Dalvad & Morten Tannert 2021. Introduction: The Role of Computational Methods within the Humanities and Social Sciences. *Scandinavian Studies in Language* 12(1), i–iv.

Bøegh, Kristoffer Friis, Peter Bakker, Inger Schoonderbeek Hansen & Carsten Levisen 2023. The Beginning of Quantitative Sociolinguistics in the Nineteenth Century: The Dane Anker Jensen (1878–1937) and his pioneering study "The Linguistic Situation in the Parish of Aaby, Aarhus County" (1898). *Historiographia Linguistic*a 49(2/3), 336–354.

Goldshtein, Yonatan & Rasmus Puggaard 2019. Overblik over danske dialektoptagelser. *Ord & Sag* 39, 18–28.

Gudiksen, Asgerd 2021. Ømålsordbogens samlinger som sproghistorisk kilde. *Danske Talesprog* 21, 11–28.

Hansen, Inger Schoonderbeek 2011. Livslange indsatser – om arbejdet hen imod Jysk Ordbog. I: Arboe, Torben & Inger Schoonderbeek Hansen (red.), *Jysk, ømål, rigsdansk mv. Studier i dansk sprog med sideblik til nordisk og tysk. Festskrift til Viggo Sørensen og Ove Rasmussen.* Århus: Peter Skautrup Centret for Jysk Dialektforskning, Nordisk Institut, Aarhus Universitet, 299–316.

Hansen, Inger Schoonderbeek 2020. Jysk Ordbog – af hvem, til hvem og hvorfor? I: Sandström, Caroline, Ulla-Maija Forsberg, Charlotta af Hällström-Reijonen, Maria Lehtonen, & Klaas Ruppe (red.), *Nordiska studier i lexikografi* 15 (Skrifter utgivna av Nordiska föreningen för lexikografi. Skrift nr 16). Helsingfors, 135–143.

Hovmark, Henrik 2011. Data og repræsentativitet i ordbogsarbejdet. I: Eaker, Birgit, Lennart Larsson & Anki Mattisson (red.), *Nordiska studier i lexikografi 11. Rapport från Konferens om lexikografi i Norden*. Lund, 296–308.

Hovmark, Henrik 2023. Seddelsamlinger – historisk arkivmateriale eller levende resurse? I: Holmer, Louise et al. (red.), *Nordiska studier i lexikografi 16*. Lund & Göteborg: Nordiska föreningen för lexikografi, 165–176.

Hovmark, Henrik & Asgerd Gudiksen 2018. Digitization of the Collection at Ømålsordbogen – the Dictionary of Danish Insular Dialects: Challenges and Opportunities. *CEUR Workshop Proceedings 2084*, 341–348.

JO = *Jysk Ordbog*. 2000–. <www.jyskordbog.dk>. Tilgået februar 2023.

Nielsen, Bent Jul 1986. Om pronominet *sin* i jysk. *Danske folkemaal* 28, 41–101.

Ordbog over det danske sprog. 1918–2005. <www.ordnet.dk/ods>. Tilgået februar 2023.

Ore, Christian Emil & Nina Kristiansen 1998. *Dokumentasjonsprosjektet. Sluttrapport 1992–1997*. Oslo: <https://www.dokpro.uio.no/sluttrapp.pdf>. Tilgået februar 2023.

Svendsen, Mette-Marie Møller 2023. Brugernes blik på Jysk Ordbog: En brugerundersøgelse i leksikografiens tegn. I: Holmer, Louise et al. (red.), *Nordiska studier i lexikografi 16*. Lund & Göteborg: Nordiska föreningen för lexikografi, 389–401.

Svensén, Bo 2004. *Handbok i lexikografi: Ordböcker och ordboksarbete i teori och praktik*. Stockholm: Norstedts Akademiska Förlag.

Sørensen, Viggo 2019. *Redaktionsregler for Jysk Ordbog i iLEX-regi*. <www.jysk.au.dk/publikationer/retningslinjer-for-jysk-ordbogs-redaktion. Tilgået februar 2023.

van Keymeulen, Jacques 2018. The Dialect Dictionary. I: Boberg, Charles, John Nerbonne & Dominic Watt (red.), *The Handbook of Dialectology*. Hoboken, NJ: John Wiley & Sons, 39–56.

Ømålsordbogen. En sproglig-saglig ordbog over dialekterne på Sjælland, Lolland-Falster, Fyn og omkringliggende øer 1992ff. Gudiksen, Asgerd, Henrik Hovmark & Karen Margrethe Pedersen (red.), Center for Dialektforskning, Institut for Nordiske Studier og Sprogvidenskab, Københavns Universitet. København: C.A. Reitzels Forlag (bd. 1–8 og tillægsbind), Odense: Syddansk Universitetsforlag (bd. 9–12).

Det Centrale Ordregister. Et indeks for det danske ordforråd – en gave til dansk sprogteknologi

Peter Juel Henrichsen

Det Centrale Ordregister (COR) is a newly published Danish lexicographic register. Each Danish lemma and each Danish word form has (or can have) a unique COR index associated with it. By design, COR thus provides a stable and authoritative reference to the Danish vocabulary at large. COR includes (*i*) morphological information for 64,000 Danish lexemes (viz. all word forms covered by the official Danish orthographic norm), and (*ii*) a linking structure making a wide range of Danish linguistic resources inter-operable (corpora, dictionaries, term banks etc.). COR is especially intended for use in the language technological sectors. COR is published open-source (under licence CC0).

NØGLEORD: sprogteknologi, dansk sprog, automatisk tekstanalyse, leksikografi til NLP, CLINK

1. Det Centrale Ordregister – bedre input til sprogteknologerne

COR (det Centrale OrdRegister) er et register over det danske ordforråd særligt udviklet til brug i sprogteknologi. Hvert dansk lemma og hver dansk ordform har – eller kan få – tilknyttet et unikt COR-indeks.

Lemma

dansk	adj	`COR.15006`

Ordform

danske	adj.sg.best	`COR.15006.302.01`
danske	adj.plur	`COR.15006.303.01`

Tekst annoteret med COR-indekser er naturligt disambigueret for homografi og er dermed egnet som input til mange slags sprogteknologi som fx stave- og grammatikkontrol, tekstanalyse, talesyntese, maskinoversæt-

telse og natursprogsgrænseflader ('natural language interfaces'). Denne artikels hovedformål er at introducere COR som leksikologisk projekt, som leksikografisk grundressource og som værktøj for udvikleren af NLP ('natural language processing'). I artiklens sidste afsnit omtales projektets organisation og tidsplan.

Betegnelsen 'COR' er en parallel til 'CPR' (Det Centrale Personregister). Hver dansker får ved fødslen et CPR-nummer som følger personen hele livet. Med sit CPR-nummer har man let adgang til at søge i samfundets databaser – også baser der ikke er relateret indbyrdes – med information om adresse, sundhed, uddannelse, skatteforhold, og så videre. I samme ånd tildeler COR et unikt og uforanderligt indeks til hvert dansk ord.

Nu er det jo ikke fornuftigt – endsige muligt! – at leksikalisere det samlede danske ordforråd, alene fordi man ikke kan trække en klar grænse mellem etablerede leksemer og kortlivede neologismer. Ikke desto mindre *kan* ethvert dansk ord (lemma og ordform) få tilknyttet et COR-indeks. Vi har søgt at opnå denne 'praktiske fuldstændighed' ved at inddele Det Centrale Ordregister i tre niveauer, hvoraf niveau 1 dækker det centrale ordforråd og udgør det leksikalske grundlag som lemmaer og ordformer i de øvrige niveauer refererer til, mens COR's niveau 2 og 3 er åbne for tilgang af nye leksemer. Denne artikel fokuserer på COR's niveau 1 (herefter 'COR_1'), altså den leksikalske grundressource som hele Det Centrale Ordregister refererer til. Denne del af COR er udviklet af Dansk Sprognævn (dsn.dk) mens Det Danske Sprog- og Litteraturselskab (dsl.dk) og Center for Sprogteknologi (cst.dk) udvikler en ordsemantisk database i COR's niv. 2 (Pedersen et al. 2022). En indføring i hvordan eksisterende ordbøger gøres COR-kompatible gives i Widmann (2023) og mere forenklet i Dideriksen et al. (2022).

1.1. COR niveau 1 – det leksikalske fundament

COR_1 er en database med omkring 64.000 danske lemmaer (500.000+ ordformer) med oplysning om ortografi, bøjningsformer og sammensætningspotentiale. Den leksikalske dækning er identisk med Retskrivningsordbogens (Schack et al. 2012; se også Link1 1997), og oplysningerne i COR_1 er officielt ratificeret (Lov om Dansk Retskrivning, Link2 1997). COR_1 dækker det almindelige danske ordforråd forstået som de lemmaer der (*i*) benyttes alment (dvs. ikke er typiske for bestemte alders- eller sam-

fundsgrupper), (*ii*) forekommer stabilt over tid (efter redaktionens vurdering) og (*iii*) generelt ikke refererer til specifikke organisationer, sagsforhold, produkter og personer. COR_1 er – ifølge sine selektionskriterier, informationstyper og status som ortografisk norm – relevant for næsten alle typer af NLP-applikationer og er dermed det naturlige fundament for COR som helhed.

TABEL 1. Et udvalg af ordformer fra COR_1: bøjning, indeks, status i RO (se afsn. 1.1) og information om lemma.

Item	*Ordform*	*Bøjning*	*COR1-indeks*	*#RO*	*Lemma*
a.	*se*	vb.inf.akt	COR.30600.200.01	1	Leksem: {se}
b.	*ses*	vb.inf.pass	COR.30600.201.01	1	Klasse: verbum
c.	*ser*	vb.præs.akt	COR.30600.203.01	1	
d.	*så*	vb.præt.akt	COR.30600.206.01	1	*(udvalgte former)*
e.	*sås*	vb.præt.pass	COR.30600.207.01	1	
f.	*så*	vb.inf.akt	COR.30901.200.01	1	Leksem: {så}
g.	*sås*	vb.inf.pass	COR.30901.201.01	1	Klasse: verbum
h.	*sår*	vb.præs.akt	COR.30901.203.01	1	*(udvalgte former)*
i.	*så*	adv	COR.10147.900.01	1	Leksem: {så} Klasse: adverbium
j.	*så*	konj	COR.00364.970.01	1	Leksem: {så} Klasse: konjunktion
k.	*dansk*	adj.sg.ubest.fk	COR.15006.300.01	1	Leksem: {dansk}
l.	*dansk*	adj.sg.ubest.itk	COR.15006.301.01	1	Klasse: adjektiv
m.	*danske*	adj.sg.best	COR.15006.302.01	1	
n.	*danske*	adj.pl	COR.15006.303.01	1	*(udvalgte former)*
o.	*danskere*	adj.kompar	COR.15006.304.01	0	
p.	*danskest*	adj.superl.sg.ubest	COR.15006.305.01	0	
q.	*danskeste*	adj.superl.sg.best	COR.15006.306.01	0	
r.	*danskeste*	adj.superl.pl	COR.15006.307.01	0	
s.	*imedens*	konj	COR.00367.970.01	1	Leksem: {imedens}
t.	*imens*	konj	COR.00367.970.02	1	Klasse: konjunktion

Tabel 1 giver eksempler på COR_1-indeksering. Bemærk i række *d.*, *f.*, *i.* og *j.* at homografen *så* indekseres forskelligt efter sin klassifikation som vb.præt.akt, vb.inf.akt, adv og konj. Række *s.* og *t.* viser hvordan det tredje numeriske felt bruges til at adskille rent ortografiske varianter (`01` og `02`), altså ordformer der ikke afviger fra hinanden i betydning, bøjning, udtale osv. (i disse tilfælde vil der ofte være en ældre og en nyere stavemåde hvoraf den ene er på vej ud af retskrivningen; dette vil dog ikke påvirke det bagvedliggende lemmas COR-indeks). I kolonnen #RO vises ordformernes RO-status: 1 betyder at staveformen er eksplicit normeret i Retskrivningsordbogen, 0 bruges til former der kun er indirekte normeret (afledt af generelle staveregler). De fleste 0-former forekommer sjældent i hverdagens tekstarter, men træffes nu og da (ofte spøgende, 'tongue in cheek'). Adjektivet *danskeste* (se tabel 1) har fx 13 forekomster i korpuset DAGW:DANAVIS svarende til 0,46 ppm (DANAVIS er en del af DAGW, et stort dansk referencekorpus; omtales nærmere i afsn. 3.5).

Ordformerne i COR_1 dækker typisk 96-99 % af teksterne (undtaget proprier, numeralier og tekniske symboler) i populære tekstgenrer såsom aviser, magasiner, wiki, studietekster på lavere niveau, skønlitteratur, altså de genrer som korpuslingvister typisk udvælger til *balanced text corpora*. COR_1 er gratis, frit tilgængeligt og åbent for enhver anvendelse. Det kan downloades som fuldformsliste fra https://ordregister.dk, hvor man også finder relevante manualer, programmer, oplysning om COR-kompatible orddatabaser osv.

I det følgende afsnit giver vi eksempler på hvordan COR-opmærkning kan føre til forbedret sprogteknologi ved at udnytte COR's grundlæggende princip om entydig leksikalsk reference. Afsnit 3 introducerer CLINK – en applikation som annoterer hvert token i en tekst med dets relevante COR-indeks. Artiklen slutter med en perspektivering og en opfordring. Dansk bør ikke være det eneste sprog med et centralt ordregister.

2. Nøgen tekst er dårligt input

Ortografien er – på trods af sin dominerende rolle som sproglig repræsentation i alle moderne samfund – på mange måder en upålidelig afbildning af et ordforråd. Dette gælder i alle sprog med eksempler på homo-

grafi, men a fortiori i dansk, som er berygtet for sine uigennemskuelige skrift-til-lyd-principper. Som alle nordboer ved, er det ret umuligt at forudse hvordan et skrevet dansk ord lyder eller hvordan et udtalt ord staves (medmindre man ved det i forvejen). Tag som eksempel tekstordet *for*, her lydskrevet i den såkaldte SAMPA-formalisme (Link3 1995).

Skriftform	Udtale	Brug (eksempel)	Ordklasse
for	[fC]	*for fanden*	*for*/præp
for	[f"C]	*for og imod*	*for*/adv
for	[f"O:]	*for og bag*	*for_og_bag*/flerordsudtryk
for	[f"oR?]	*hun for afsted*	*for*/vb
for	[f"o:?R]	*frakkens for er slidt*	*for*/sb

Det er altså ikke muligt for fx en talesyntese at realisere en (dansk) tekst som oplæst tale uden at tekstordene er blevet leksikalsk disambigueret. På samme måde vil en automatisk oversætter gå i knæ over et input fuld af homografer.

Maj så Find så alle frøene, så hun bestilte nye så snart de slap op

Afprøv denne tekst som input til Google Translate, og få et underholdende svar.

2.1. PoS-annoteret tekst

En klassisk og velafprøvet måde at disambiguere en tekst på er ved at PoS-tagge den, altså give hvert ord et ordklassemærke (et *tag*).

(*a'*)	*Maj*/prop	*så*/vb	*Find*/prop	*så*/vb	*frø*/sb
(*b'*)	*Maj*/fornavn	*så*/vb.præt.akt	*Find*/fornavn	*så*/vb.inf.akt	*frø*/sb.itk.pl.ubest

PoS-tagging kan gøres manuelt eller overlades til en automatisk PoS-tagger (programmer af denne type kaldes 'classifiers'). Mange PoS-taggere er kun trænet til at fordele input i de overordende ordklasser, verbum, substantiv, adjektiv, konjunktion osv. Denne analysedybde er tilstrækkelig til en del formål. Den fanger ikke den lemmatiske forskel på de to instanser af *så* i eksempel *a'*, men dette vil ikke påvirke kvaliteten af

en talesyntese da de to instanser udtales ens; tilsvarende for homografen *frø* der både kan være en neutrums- og en utrumsform, men med samme udtale. En oversætter har naturligvis brug for den finere klassifikation i *b'*.

I andre situationer er disambiguering med PoS-tagging imidlertid utilstrækkelig eller irrelevant uanset den morfosyntaktiske finhedsgrad. Et nominalsyntagme som *flot fyr* kan ikke disambigueres leksikalsk ved PoS-analyse alene.

	flot	*fyr*	
(*c'*)	adj.sg.ubest.itk	sb.itk.sg.ubest	≈ NICE STOVE
(*d'*)	adj.sg.ubest.fk	sb.fk.sg.ubest	≈ HANDSOME FELLOW
(*e'*)	adj.sg.ubest.fk	sb.fk.sg.ubest	≈ BEAUTIFUL PINE TREE

Som det ses, er de to eksempler *d'* og *e'* formelt uskelnelige. Det er dårligt nyt for oversætteren, men her også for talesyntesen idet *fyr fellow* og *fyr pine* udtales forskelligt. For at redde situationen kunne vi overveje at tilføje nogle træksemantiske tags til PoS-inventaret. Men hvilke?

Det har vist sig vanskeligt at designe et versatilt tagsæt som er lige relevant til alle (sprogteknologiske) formål. Jagten på den universelle tagger er endt uden resultat og har – trods fem årtiers datalingvistisk indsats – efterladt os med et pletora af tekstkorpora der ikke kan lægges sammen på grund af forskelle i annotationsart og annotationsdybde. I vores undersøgelse (Kirchmeier et al. 2019, Kirchmeier et al. 2020) kunne vi konkludere at, ud af 100+ store danske korpora skabt for offentlige midler gennem de seneste 30 år, kun cirka fem har relevans i dag. Resten er endt på datakirkegården, ikke mindst på grund af inkompatible annotationsformater.

2.2. Leksikalsk disambigueret tekst

Oprettelsen af Det Centrale Ordregister giver mulighed for at skifte perspektiv. Vi foreslår at man erstatter ideen om en tagger med ideen om en linker. Linkeren klassificerer ikke tekstens elementer i forhold til et forud givet (og projektrelateret) inventar af kategorier, men annoterer i stedet hvert tekstord med en reference til et centralt ordregister (in casu COR_1) som i sig selv er applikationsneutralt. Sammenlign den lek-

sikalsk disambiguerede tekst i *c"-e"* herunder med den morfosyntaktisk disambiguerede tekst i *c'-e'* (bemærk især adskillelsen af *fyr*-instanserne i *d"* og *e"*).

	flot	fyr	
(*c"*)	COR.17043.301.01	COR.77168.120.01	≈ NICE STOVE
(*d"*)	COR.17043.300.01	COR.53883.110.01	≈ HANDSOME FELLOW
(*e"*)	COR.17043.300.01	COR.84448.110.01	≈ BEAUTIFUL PINE TREE

Princippet om COR-linking kalder naturligvis på to spørgsmål: Hvad gør man med tekstord som ikke forekommer i COR_1? Og hvad med de applikationer som behøver leksikalsk information der ikke findes i COR_1 (udtale, semantik, terminologiske relationer osv.)? Disse spørgsmål bliver belyst i det følgende.

3. CLINK ver. 1.0

Dansk Sprognævn udgav d. 1/10 2022 en COR-linker, kaldet CLINK, som frit kan downloades, bruges og videreudvikles (under licens CC0). CLINK er en input-output-applikation; den læser en tokenstreng (*plain text*) og udskriver den samme streng med hvert token annoteret med COR_1-indeks. Ord som ikke forekommer i COR_1, tagges med <OOV> (Out Of Vocabulary).

Algoritmen i CLINK 1.0 bygger på tre forskellige strategier; de præsenteres i 3.1-3.3 herunder, og interaktionen imellem dem i 3.4. Læsere uden interesse for datalingvistik kan springe til afsnit 3.5 uden skade for sammenhængen.

3.1. Strategi 1: Window-one

Den simpleste linker-strategi ser kun på ét token ad gangen. I parserteori omtales denne strategi somme tider som 'window-1-decisions', altså de beslutninger der kan træffes om et token uden at skele til dets omgivelser. I nogle situationer er denne strategi optimal, nemlig for former med nul eller ét indeks i COR_1 (her har man ingen gavn af konteksten). Men når et token er homografisk, er strategi 1 (S1) naturligvis dårlig (her afhænger den gode beslutning netop af de syntaktiske omgivelser).

Formelt betragtet er S1 komplet, forstået sådan at den altid kan annotere alle tokens i input (enten med et COR_1-indeks eller <OOV>). Ingen af de andre strategier er komplette i denne forstand, og S1 er derfor nødvendig som fallback. I classifier-termer har S1 en *recall* på 1.0. For de tokens som ikke er homografer i COR_1, er *precision* også 1.0, men for de homografe tokens er præcisionen dårlig: Her linker S1 jo bare til den mest frekvente form i COR_1 (hvis den har adgang til en frekvenstabel) eller også må den træffe et tilfældigt valg – hvad skulle den ellers gøre?

3.2. Strategi 2: De lokale omgivelser

Strategi 2 (S2) træffer beslutninger baseret på de morfosyntaktiske omgivelser omkring et token. Det simpleste eksempel er flerordsforbindelsen (MWE, 'multi word expression'), hvor et token indgår i et *n*-gram som er leksikaliseret. Der er ikke stor forskel, hvad den syntaktiske funktion angår, på en MWE og et enkelt leksem, og ofte er MWE'er leksikaliseret i COR_1 som rene ortografiske varianter.

ingen ting	pron	`COR.00561.991.01`
ingenting	pron	`COR.00561.991.02`
gud hjælpe mig	adv	`COR.12501.900.01`
gudhjælpemig	adv	`COR.12501.900.02`

Da CLINK arbejder token-for-token, er hvert token i en MWE i linkerens output annoteret med et link.

Der hvor S2 er mest effektiv, er til analyse af nominalsyntagmer hvor morfologisk kongruens spiller en central rolle (især konstruktioner som DET ADJ* CN MOD*, hvor DET er et determinativled, ADJ* en gruppe adjektiver, CN et appellativ og MOD* pladsen for præpositionsforbindelser og andre postmodifiers). Her gemmer sig den største (og interessanteste) udfordring for linkerudvikleren; af pladsgrunde må vi dog her nøjes med at illustrere S2 med nogle eksempler på leksikalsk disambiguering ved hjælp af unifikation af morfologiske træk.

S2 benytter den såkaldte PAROLE-formalisme til morfologisk trækstruktur (Keson 1999, Henrichsen 2002). En trækstruktur som *sb.itk. pl.ubest.gen* i COR_1 (svarende til ordformer som *bogstavers* og *æblers*)

ser i PAROLE sådan ud: NCNPG==I. Det første tegn er hovedordklassen, og hvert efterfølgende tegn koder for et bestemt morfologisk træk. Tagget NCNPG==I svarer således til Noun-Commonnoun-Neuter-Plural-Genitive-*void*-*void*-Indefinite. En trækværdi der er uspecificeret, markeres med et punktum, mens et træk der er udefineret markeres med = eller –. PAROLE-formalismen er udviklet specielt til brug i taggere (og linkere).

et stort fyr	P**I**–NSU---	ANPNSU=**I**U	NCNSU==**I**	≈ A BIG STOVE
et flot fyr	P**I**–NSU---	ANP.SU=**I**U	NCNSU==**I**	≈ A NICE STOVE
fyrre orange fyr	ACP--U=--	ANP...=.U	NCNPU==I	≈ FORTY ORANGE STOVES
fyrre flotte fyrre	ACP--U=--	ANP.PU=.U	NCCPU==I	≈ FORTY BEAUTIFUL PINES

I eksemplerne er kongruenser i trækket *bestemthed* markeret med **fed**, trækket *numerus* med understregning og trækket *genus* med dobbeltunderstregning. CLINK's opgave er altså at udvælge de COR_1-links der får hele kongruensregnestykket til at gå op. I datalogien kaldes denne øvelse 'unification'. Læsere med programmeringserfaring bliver næppe overraskede over at koden til S2 er skrevet i PROLOG, et sprog som er særligt egnet til netop unifikation.

3.3. Strategi 3: Den videre kontekst

S1 og S2 supplerer hinanden fint. S1 garanterer at linkeren som sådan har en *recall* på 1.0 mens S2 bidrager til en stærkt forbedret *precision*. Men der er stadig 'mørke områder' i en almindelig tekststreng med homografi der er uden for begge strategiers rækkevidde.

en flot fyr	*fyr* / sb.fk.sg.ubest / COR.53883.110.01
en flot fyr	*fyr* / sb.fk.sg.ubest / COR.84448.110.01

Godt nok kunne vi slå fast i afsnit 2.2 at COR-links (i modsætning til PoS-tags) kan adskille eksemplets to instanser af *fyr* leksikalsk. Men hvordan skal CLINK vælge det relevante link? Hvis input til linkeren kun består af *en flot fyr*, er afgørelsen umulig (selv for et menneske); men hvis input er en længere tekst, og *fyr* altså har en kontekst der går forud, kan denne afsøges for leksikalske pejlemærker. Dette er Strategi

3's funktion (S3). Den har adgang til en associationstabel som fx knytter ordformen

fyr	COR.84448.110.01	PINE TREE

til en række af semantisk associerede lemmaer som

træ	sb	COR.44241
nåletræ	sb	COR.91480
brænde	sb	COR.91213
brænde	vb	COR.30160

(og tilsvarende for *fyr* FELLOW). Da CLINK jo, for hvert token i inputstrengen, har adgang til sin egen analyse af de forudgående tokens, kan den (med lidt held) bruge de associerede lemmaer i venstrekonteksten som indicier. CLINK version 1.0 har kun omkring 80 leksikalske indgange i sin associationstabel – men vi forventer en markant udvikling på denne front når COR-projektet i slutningen af 2023 offentliggør sin omfattende ordsemantiske database (se Pedersen et al. 2022).

3.4. Parallel versus seriel kobling

Der er (grundlæggende) to måder at koble de tre strategier på, enten parallelt eller i serie. Ved parallelkobling, også kaldet 'voting', prøves alle tre strategier på ethvert token – og den endelige beslutning om inputordets link træffes ved afvejning af de tre bud. Hvis S1 er meget sikker i sin vurdering, mens S2 slet intet bud har, og S3 er usikker, går beslutningen til S1. Generelt træffes valget altså i forhandling mellem strategierne, og hvis de ikke kan blive enige, sørger en dommerstrategi for at lægge valget hos den strategi som viser størst sikkerhed ('confidence'). Parallelstrategier kan nå meget høje succesrater, men er vanskelige at optimere på grund af de mange variabler som programmøren skal justere.

Den serielle kobling fungerer som et samlebånd hvor hver strategi anvendes efter tur. Den har færre frihedsgrader og er lettere et optimere, og vi har derfor foretrukket en seriekoblet algoritme til CLINK ver. 1.0.

TOKEN → **S1**$^{-}$ → **S2** → **S3** → **S1**$^{+}$ → LINK

Bemærk at S1 i første omgang kun får myndighed til at beslutte links for ikke-homografe inputtokens (hvor *precision* er 1.0). Kun hvis hverken S2 eller S3 kan beslutte sig for et link, får S1 det sidste ord (men med lav *precision*).

Vi deler koden til S1, S2 og S3 med alle interesserede, og vores håb er at andre vil tage udfordringen op så vi i fællesskab får skabt en hurtigere, mere præcis COR-linker til glæde for alle.

3.5. Nyeste CLINK-udvikling

Dansk Sprognævn arbejder i disse måneder på at COR-linke korpus DAGW (Danish Gigaword Corpus, se www.sprogteknologi.dk). DAGW består af uannoterede tekster fri for copyright. Korpusteksterne er valgt i genrer som de fleste danskere er i jævnlig kontakt med, nærmere motiveret i Derczynski et al. (2021). Teksterne er uannoterede og foreligger som 'plain text' (i formatet UTF-8). DAGW har, trods sin unge alder, vundet stor udbredelse som dansk referencekorpus til både forskning og udvikling. Dansk Sprognævn sigter mod at have en komplet COR-linket DAGW klar inden jul 2023. Først på dét tidspunkt bliver det muligt at udmåle CLINK's performans kvantitativt. I skrivende stund har vi foretaget kvalitative målinger (stikprøvebaserede, manuelt evaluerede); efter vores bedste skøn har CLINK 1.0 en fejlrate (målt på blandede tekstgenrer a la DAGW) på 2.5-4.5 %. Læs mere om COR-linking i Henrichsen (2023).

4. Videre perspektiver

4.1. Kontrollerede udvidelser af Det Centrale Ordregister

Der er naturligvis mange danske ord som ikke findes i COR_1: fagord, neologismer, ældre ord, proprier og en ubegrænset mængde komposita. Desuden har mange NLP-applikationer brug for leksikalsk information som COR_1 ikke rummer.

Talesyntese	*fonetiske og prosodiske data*
Maskinoversættelse	*semantiske ækvivalenter i målsproget*
Terminologisystemer	*begreber og begrebsrelationer*
Betydningsordbøger	*ordbetydninger*
AI-baserede dialogsystemer	*sætnings- og diskurssemantik, verdensviden, ...*

Både nye leksemer og supplerende leksikalske data til eksisterende leksemer kan, på systematisk måde, gøres tilgængelige via Det Centrale Ordregister. Dertil har vi COR's niveauer 2 og 3 med eksplicit leksikalsk linking til COR_1 (Widmann 2023, Dideriksen et at 2022). Sprogressourcer som overholder COR-formatet, kaldes *COR-kompatible*.

Vi opfordrer alle danske korpus- og ordbogsredaktioner til at gøre deres ressourcer COR-kompatible (manualer findes i www.ordregister.dk). Øvelsen består i (*i*) at COR-indeksere hver leksikalsk indgang og/eller hvert tekstord, (*ii*) at udarbejde en afbildningstabel mellem de nye indekser og COR_1. Man kan vælge enten at publicere sin COR-kompatible ressource eller at nøjes med at udgive afbildningstabellen og så bevare sit indhold bag fx en betalingsmur. I begge tilfælde øger man sin ressources værdi som data til sprogteknologiske anvendelser.

4.2. Projekt COR – nu og i fremtiden

Det Centrale Ordregister er, i skrivende stund, stadig under udvikling. Projektgruppen består af Dansk Sprognævn (design af det formelle rammeværk, udvikling af COR_1), Det Danske Sprog- og Litteraturselskab (leksikografiske udvidelser og ordsemantisk annotation) og Center for Sprogteknologi (maskinlæring i forbindelse med semantisk annotation). Projektarbejdet er støttet af en treårig bevilling fra Innovationsministeriet, administreret af Digitaliseringsstyrelsen (digst.dk); læs mere om den danske sprogteknologiske satsning i Kirchmeier et al. (2019) – og besøg også www.sprogteknologi.dk, hvor den danske sprogteknologiske satsning er omtalt i detaljer.

Når COR-bevillingen rinder ud med udgangen af 2023, er det vigtigt at de nyudviklede sprogressourcer fortsat vedligeholdes og udvikles. Dansk Sprognævn, som har en naturlig forpligtelse til at støtte alle aspekter af dansk sprogbrug – inklusive de sprogteknologiske – har derfor indvilget i at stå for den fremtidige administration af Det Centrale Ordregister. Opgaven består ikke kun i at vedligeholde COR_1-ressourcens tilgængelighed og aktualitet, men også i at vejlede om forberedelsen af nye COR-kompatible ressourcer.

4.3. Nordisk samarbejde – et CALL for COR

Færøerne har nu udviklet sin egen parallel til COR kaldet OTAL (Simonsen et al. 2022). Det er en imponerende bedrift, og Færøernes eksempel viser at hvor der er vilje, er der vej. Vi vil hermed opfordre alle de nordiske sprogsamfund, større såvel som mindre, til at oprette deres egne centraliserede ordregistre til gavn for sprogteknologien i hele Norden.

Referencer

Litteratur

Derczynski, Leon et al. 2021. The Danish Gigaword Corpus. *Proceed. of NODALIDA-23*. Linköping Electronic Conference Proceedings 178 (2021).

Dideriksen, Christina, Peter Juel Henrichsen & Thomas Widmann 2022. Det Centrale Ordregister. I: *Nyt Fra Sprognævnet*. Oktober 2022. ISSN 2446-3124.

Henrichsen, Peter Juel 2002. *Sidste Års Aviser*. I: Institut For Datalingvistik (KU): LAMBDA 27.

Henrichsen, Peter Juel 2023. Diktaroriske Beføelser. Om Ord og Uord i Det Centrale Ordregister. I: *Proceedings of MUDS19* (Møde om Udforskningen af Dansk Sprog). Aarhus Universitet.

Keson, Britt 1999. *Vejledning til det Danske Morfosyntaktisk Taggede PAROLE-korpus*. DSL Press.

Kirchmeier, Sabine, Peter Juel Henrichsen & Philip Diderichsen 2019. *Dansk Sprogteknologi i Verdensklasse*. Rapport fra sprogteknologiudvalget under Dansk Sprognævn nedsat af Kulturministeriet. ISBN 978-87-89410-77-7.

Kirchmeier, Sabine, Bolette Sandford Petersen, Peter Juel Henrichsen; Sanni Nimb & Philip Diderichsen 2020. World Class Language Technology – Developing a Language Technology Strategy for Danish. *Proceedings of LREC 2020*.

Pedersen, Bolette, Nathalie Carmen Hau Sørensen, Sanni Nimb, Sussi Olsen, Ida Flørke & Thomas Troelsgård 2022. Compiling a Suitable Level of Sense Granularity in a Lexicon for AI Purposes: The Open Source COR Lexicon. *Proceed. of LREC2022*.

Schack, Jørgen et al. (red.) 2012. *Retskrivningsordbogen*. 4. udgave. Dansk Sprognævn.

Simonsen, Annika, Sandra Saxov Lamhauge, Iben Nyholm Debess & Peter Juel Henrichsen 2022. Creating a basic language resource kit for Faroese. *Proceed. of LREC2022*.

Widmann, Thomas 2023. Det Centrale Ordregister og dets leksikografiske anvendelser. I: Holmer, Louise et al. (red.), *Nordiska studier i lexikografi 16*. Lund & Göteborg: Nordiska föreningen för lexikografi, 415–430.

Links (verificeret maj 2023)

Link1 (1997) Lov om Dansk Sprognævn (LOV nr 320 af 14/05/1997)
https://www.retsinformation.dk/eli/lta/1997/320
https://kum.dk/ministeriet/organisation-og-institutioner/bestyrelser-raad-naevn-og-udvalg/dansk-sprognaevn-repraesentantskab

Link2 (1997) Lov om dansk retskrivning (LOV nr 332 af 14/05/1997)
https://www.retsinformation.dk/eli/lta/1997/332

Link3 (1995) Dansk SAMPA, fonetisk alfabet til computerbrug (se også DSN's modificerede SAMPA i www.dsn.dk)
https://www.phon.ucl.ac.uk/home/sampa/danish.htm

Pragmatiska markörer i samtal – en webbaserad ordbok för isländskt talspråk

Helga Hilmisdóttir

In this article, I present the first draft of a web-based dictionary for Icelandic, the *Samtalsorðabók* 'Conversational dictionary'. The lemmas consist of pragmatic markers that are used in talk-in-interaction, i.e. words and phrases that have a pragmatic- or discourse structuring-function. These include words such as discourse particles, discourse markers, greetings, address terms and swearwords. The goal of this dictionary is to give spoken interaction its own platform and thus to raise awareness and knowledge about the peculiarities of Icelandic conversational language and to show recurring patterns. In the first version, *Samtalsorðabók* contains 171 lemmas. In addition to a short definition which focuses on discourse functions, each lemma is presented with at least one example of use, i.e. a short segment from an authentic conversation that contains the word or phrase in question. The user can listen to the sound clip, read a transcription on the screen and click on a set of keywords that refer to different discourse functions or contexts. The keywords function as tags that give the user a list of lemmas associated with the same features.

NYCKELORD: webbordbok, samtal, pragmatiska markörer, samtalsanalys

1. Inledning

I denna artikel presenteras en ny webbordbok, *Samtalsorðabók*, som fokuserar på pragmatiska markörer i isländskt samtalsspråk. Uttrycket *pragmatiska markörer* används i denna artikel som en övergripande term för ord och fraser som har det gemensamt att de hör till det pragmatiska eller diskursstrukturella planet, t.ex. diskurspartiklar, diskursmarkörer, tilltal, hälsningsfraser, svordomar och artighetsmarkörer. Som viktig förebild för ordboken fungerar pilotprojektet *Ordbog over Dansk Talesprog* som utarbetades under en tid vid Köpenhamns universitet och fokuserade på interjektioner (Hansen 2015).[1]

1 *Ordbog over Dansk Talesprog* innehåller 50 olika uppslagsord som belyses med korta ljudklipp och transkriptioner. Arbetet har inte fortsatt efter att pilotprojektet slutfördes.

För att svara på frågan *varför* det finns ett behov för en samtalsordbok för isländska vill jag lyfta fram en insändare som en ledamot på Alltinget skrev i en av Islands största dagstidningar. I texten uttrycker han en stark negativ attityd mot ett etablerat talspråkligt drag i isländska, nämligen diskurspartikeln *þúveist* 'du vet': "ministrar deltar i teveprogram och tränger in 'du vet', antagligen för att deras tankar är kaotiska, oftare än man kan hålla reda på. [...] Jag frågar mig ofta, vad fan? Varför berätta något jag redan vet?" (Bjarnason 2020; min översättning; se Hilmisdóttir 2001 om *þúveist*).

Ledamoten är inte ensam om att använda oplanerat talspråk som en källa till kritik. I sociala medier förs det regelbundet diskussioner kring personer som har uttryckt sig i medierna och ofta handlar kritiken om användningen av pragmatiska markörer. I denna insändare väljer ledamoten att tolka *þúveist* 'du vet' ordagrant och låtsas som om han inte vet hur diskurspartikeln används i samtal. Partikeln *þúveist* finns inte heller i isländska ordböcker (t.ex. *Íslensk nútímamálsorðabók*). Precis som i andra språk bygger isländska ordböcker i första hand på skriftliga källor, i synnerhet tryckt material (se diskussion i Svavarsdóttir 2007). Information om talspråkliga drag förekommer däremot sporadiskt och ofta tas de inte in i ordböckerna förrän de börjar dyka upp i skriftspråk, t.ex. i litterära dialoger. Vidare kan man påpeka att i traditionella ordböcker framställs talspråket ofta som en avvikelse eller stilvariant i förhållande till skriftspråket (se *Ordbog over Dansk Talesprogs* hemsida). Genom att utveckla en ordbok som lyfter fram talspråkliga drag synliggörs de och dokumenteras.

Syftet med denna artikel är att presentera *Samtalsorðabók*, som i denna artikel även kallas *samtalsordboken*, och diskutera idén bakom den. Artikelns disposition är som följer: I avsnitt 2 diskuteras ramarna för ordboken. I avsnitt 3 beskrivs de pragmatiska markörernas framställning i traditionella, enspråkiga ordböcker och därefter, i avsnitt 4, dryftar jag teoretiska och metodologiska utgångspunkter för ordboken. I avsnitt 5 berättar jag kort om materialet som används som grund för ordboken och hur det transkriberas. I avsnitt 6 redogör jag för webbsidans innehåll, dvs. vilken typ av information som finns för varje uppslagsord, och i avsnitt 7 avrundar jag diskussionen.

2. Ramarna för *Samtalsorðabók*: syfte, målgrupp och tillvägagångssätt

Målet med *Samtalsorðabók* är tredelat:

1. att ge samtalsspråket en egen plattform och erkänna dess existens och säregenheter,
2. att visa att pragmatiska markörer fyller en viktig funktion i mänsklig interaktion och att det finns en systematik bakom användningen av dem,
3. att dokumentera ordanvändningen i olika typer av samtal där informanterna är i olika åldrar, kommer från olika delar av landet och tillhör olika sociala kategorier.

Målgruppen är stor och inkluderar bl.a. lingvister, språkintresserade modersmålstalare, andraspråkstalare på olika nivåer och översättare som arbetar med litterära dialoger eller audiovisuellt material.

Grunden för *Samtalsorðabók* lades sommaren 2021 när två studeranden vid Islands universitet, Ása Bergný Tómasdóttir och Kristín Björg Björnsdóttir, anställdes i tre månader för att bygga upp en databas med ljudklipp och transkriberade belägg. Arbetet finansierades av Studenternas innovationsfond som förvaltas av det nationella forskningscentret Rannís. Projektet leddes av Helga Hilmisdóttir, databasen och webbsidan programmerades av Trausti Dagsson och som rådgivare fungerade Þóra Björk Hjartardóttir.

Ordboken är under arbete men redan publik på webben. Den är så uppbyggd att alla uppslagsord förses med minst ett samtalsutdrag där ordet förekommer i en autentisk kontext, dvs. en sekvens som kan uppfattas som en helhet med början och slut. Det är viktigt att samtalsutdragen i ordboken varken är för korta eller för långa, och att de innehållsmässigt inte är svåra att förstå när de visas som självständiga snuttar utan den omgivande samtalskontexten. Samtalsordboken består i skrivande stund (februari 2023) av 171 uppslagsord som belyses med 403 transkriberade samtalsutdrag och ljudklipp.

Vid val av uppslagsord har vi utgått från samtalen och fokuserat på pragmatiska markörer som finns i det material som vi har valt att inkludera. Projektarbetarna samlade in inspelningar och lyssnade och valde samtals-

utdrag som innehöll belägg på pragmatiska markörer. Ljudfilerna klipptes och redigerades i ljudprogrammet Audacity, t.ex. genom att ta bort personnamn. Snuttarna matades sedan in i en databas. Huvudredaktören för samtalsordboken valde sedan representativa exempel ur databasen, placerade beläggen i en logisk ordning, formulerade definitioner och valde nyckelord.

3. Pragmatiska markörer och deras framställning i isländska ordböcker

Ur lexikografisk synvinkel består de pragmatiska markörerna av en mycket heterogen grupp ord. En del markörer har i isländska ordböcker klassificerats som adverb (*jú* 'ju', *nú* 'nu') eller interjektioner (*hæ* 'hej') medan andra har hänförts till ordklasser som substantiv (tilltalet *ástin* 'älskling'), adjektiv (tilltalet *maður* 'man', hälsningsfrasen *sæll* 'säll' och svordomen *helvítis* 'djävla') och verb (*fyrirgefðu* 'förlåt') (*Íslensk orðabók* 2002, *Íslensk nútímamálsorðabók* 2022). En del markörer består även av fler ord som t.ex. *guð minn góður* 'herre gud' och *guði sé lof* 'gudskelov', som båda anges som fasta fraser under substantivet *guð* 'gud', men som varken förses med definitioner eller anvisningar om bruk i isländska ordböcker.

Gemensamt för samtliga markörer är att förklaringar i ordböckerna är kortfattade och ofta i form av ord som ligger semantiskt nära. Till exempel placeras artighetsmarkören *fyrirgefðu* 'förlåt' under verbet *fyrirgefa* 'förlåta' och förklaras som imperativ i betydelsen *afsakaðu* 'ursäkta' (*Íslensk orðabók* 2002). I motsats till uppslagsord med semantiskt innehåll, där man kan utgå ifrån att det finns en definition, verkar ord och fraser som fyller pragmatiska funktioner ofta lämnas utan kommentar.

Vidare kan man påpeka att en del pragmatiska markörer helt och hållet saknas i ordböckerna. Delvis beror det på att ord och fraser som fyller funktioner på det pragmatiska planet snabbt kan komma in i talspråket medan det tar en längre tid för dem att etablera sig i skrift. Till en början kan markörerna ha en begränsad spridning, t.ex. enbart inom ett visst område eller i en viss åldersgrupp. De kan sedan antingen få en större spridning eller försvinna. Vid redigeringen av traditionella ordböcker är det däremot viktigt att ha ett brett tidsperspektiv, vilket betyder att det kan ta lång tid innan nya fenomen kan tas upp i en utgiven ordbok. För en språkinlärare som vill förstå talspråket här och nu kan det upplevas som en för lång process.

Men även ord och fraser som har funnits länge i talspråk och informellt skriftspråk kan saknas i ordböcker. I en undersökning av isländska talspråkskorpusar visar Hilmisdóttir (2021a:84) att endast 67% av de diskurspartiklar som förekommer i materialet finns som egna uppslagsord i isländska ordböcker (dvs. *Íslensk orðabók* och *Íslensk nútímamálsorðabók*). Det är i synnerhet diskurspartiklar som har sitt ursprung i verb och fraser av olika slag som inte inkluderas som egna uppslagsord och där den pragmatiska funktionen inte anges (t.ex. *er það* 'på riktigt', *skilurðu* 'förstår du', *fattarðu* 'fattar du', ég *meina* 'jag menar'). Pragmatiska lån från engelskan, som har blivit allt vanligare i informellt talspråk, saknas också i traditionella ordböcker (t.ex *what*, *sorry, please, oh my god*) (se Hilmisdóttir 2021b; Andersen 2014 om pragmatiska lån i norska).

I en ordbok om pragmatiska markörer kan den traditionella lexikografiska ordklassindelningen vara irrelevant för språkanvändaren och till och med vilseledande. För en språkinlärare som vill veta hur man använder *þúveist* 'du vet' hjälper det t.ex. inte att hänvisas till verbet *vita* 'veta'. Diskurspartiklar har ingen verbböjning och är syntaktiskt fria, dvs. de kan förekomma på olika platser inom en syntaktiskt organiserad tur (jfr Hilmisdóttir 2001 om *skiluru* 'förstår du' som partikel). De uppför sig på ett helt annat sätt en den ursprungliga verbfrasen. Däremot är markörernas sekventiella placering och deras prosodiska drag avgörande för deras funktion (Hilmisdóttir 2007).

4. Teoretiska och metodologiska utgångspunkter för ordboken

I en artikel om diskurspartiklar i isländskt talspråk och deras framställning i enspråkiga ordböcker formulerade jag riktlinjer för hur det lexikografiska arbetet skulle kunna gå vidare (Hilmisdóttir 2021a:97–98). Där framhävde jag vikten av att den lexikografiska beskrivningen av talspråkstypiska drag som diskurspartiklar borde utgå ifrån interaktionella funktioner, dvs. vad partiklarna gör i samtalet och inte vad det urprungliga semantiska innehållet en gång har varit. Vidare föreslog jag att man, vid val av samtalsutdrag, borde ta strukturella aspekter i beaktande. Det är viktigt att utdragen uppvisar variation angående diskurspartikelns prosodiska drag och placering i yttrandet (Hilmisdóttir 2021a:97).

Slutsatsen som jag drog av min undersökning av diskurspartiklar i talspråkskorpusar var att en möjlig lösning skulle vara ”att utveckla en egen talspråksordbok där man kunde utnyttja de digitala ordböckernas fördelar och både visa längre samtalsutdrag och använda autentiska ljudinspelningar [...]” (Hilmisdóttir 2021a:98). I och med den digitala tekniken är frågan om utrymme inte lika stor som förr. I traditionella ordböcker, i synnerhet i tryckta ordböcker där det råder en brist på utrymme, är det viktigt att både redaktionella exempel och autentiska språkprov är korta, helst inte mer än en till två repliker. Ett samtalsutdrag på 8–12 rader är för mycket för en tryckt ordbok, men på webben är situationen en annan. Webbens multimodala karaktär är det också naturligt att utnyttja. Det är enkelt att ge användarna tillgång till ljudinspelningar och lägga ut interna länkar mellan olika uppslagsord. En samtalsordbok som byggs upp på talspråkets villkor skulle kunna fungera som en annorlunda lexikografisk resurs som inte behöver följa de traditioner som har byggs upp inom den klassiska lexikografin.

Den teoretiska grundidén bakom samtalsordboken, dvs. att definiera pragmatiska markörer utifrån en sekventiell analys av autentiska samtal, bygger på forskning som gjorts inom CA-inriktad samtalsanalys och den interaktionella lingvistiken (t.ex. Steensig 2001; Lindström 2008; Couper-Kuhlen & Selting 2018), i synnerhet forskning som utgår ifrån lexikala enheter som diskurspartiklar (Hilmisdóttir 2007, 2016). Ord skapas, formas och förändras i samtal och för att förstå deras funktioner måste vi analysera dem i sin naturliga kontext, i autentiska samtal.

Uppslagsordens funktion beskrivs utgående från den handling de förekommer i, precis som man har strävat efter i *Ordbog over Dansk Talesprog* (jfr Opsahl 2015:145). Än så länge har vi inom projektet inte hunnit långt med analysen av funktioner men tanken är att skapa en lista över *nyckelord* (se diskussion om ”funktioner” i Hansen 2015). Nyckelorden omfattar samtalsanalytiska termer som antingen identifierar vad de pragmatiska markörerna gör i samtalet eller säger någonting om deras formella egenskaper, t.ex. intonation eller placering. Nyckelorden i *Samtalsorðabók* motsvarar endast delvis de kategorier som förekommer i *Ordbog over Dansk Talesprog* (som enbart fokuserar på funktioner). Tabell 1 visar en översikt över några nyckelord.

TABELL 1. Exempel på nyckelord i *Samtalsorðabók*.

funktioner	**form**
UPPBACKNING	i början av turer
KVITTERING	i slutet av turer
FORTSÄTTNINGSSIGNAL	i början av en flerledad tur
MOTTAGANDE AV INFORMATION	ensamstående
VÄRDERING	med stigande ton
PLANERINGSMARKÖR	med utdragen vokal
MEDHÅLL	

På webbsidan fungerar nyckelorden som index eller interna länkar. När ordboksanvändarna klickar på ett nyckelord får de upp en lista på uppslagsord som är indexerade med samma nyckelord. Om man t.ex. klickar på nyckelordet MEDHÅLL får man upp följande uppslagsord: *akkúrat* 'precis', *algjörlega* 'helt och hållet', *einmitt* 'mitt i prick', *ekkert smá* 'inte lite', *ekki spurning* 'ingen fråga', *heldur betur* 'minsann', *hundrað prósent* 'hundra procent', *nákvæmlega* 'precis', *same* och *true*. Listan visar att nyckelorden kan ge intressanta resultat när man vill jämföra olika markörer med varandra. Detta kan t.ex. vara en nyttig funktion för språkinlärare som vill utöka sitt ordförråd.

5. Material, transkribering och etiska frågor

Materialet som används som grund för samtalsordboken består av olika typer av samtal: inspelningar från radion, podcaster, styrda gruppsamtal och vardagliga samtal. Under den första fasen har vi huvudsakligen satsat på material från massmedier som finns ute på webben men tanken är att utöka materialet primärt med utdrag ur autentiska, vardagliga samtal.

Samtalsutdrag som används i ordboken transkriberas så att användaren kan både lyssna på ljudklippet och se transkriptionen på skärmen. Inom CA-inriktad samtalsforskning används detaljerade transkriptioner: pauser mäts med millimeters precision, skratt skrivs ut så ljudnära som möjligt och ibland används en modifierad, ljudnära ortografi som delvis avviker från vanligt skriftstandard (se t.ex. Steensig 2001:32–37). I en ordbok som riktas till bl.a. andraspråkstalare är det dock viktigt att transkriptionerna inte blir för komplicerade. Enligt Hansen (2015) används

i *Ordbog over Dansk Talesprog* skriftspråksnära ortografi och symboler används sparsamt. Tabell 2 visar de symboler som vi inkluderar i den isländska samtalsordboken.

TABELL 2. Transkriberingsnyckel för *Samtalsorðabók*.

Symbol	**Betydelse**
(.)	mikropaus
(..)	paus
(...)	uppehåll, mer än 2 sekunder
[já]	överlappningar
hh	utandning
.hh	inandning
#e:::#	knarrande röst
já	skratt
@já@	med förställd röst
.mt	smackande ljud

Som tabell 2 visar använder vi tio olika symboler som har varit viktiga för analysen. Förutom pauser och överlappande tal markerar vi även ord uttalade med knarrande röst, förställd röst, skratt och inandning. De fenomen vi har valt att skriva in i transkriptionerna anser vi vara centrala för analysen av pragmatiska markörer. Knarrande röst och smackande ljud används t.ex. på ett systematiskt sätt i början av turer och i reparationer. Eftersom det inte är önskvärt att transkriptionerna blir för komplicerade inkluderar vi inte symboler för falsett eller viskande röst som används mer sporadiskt och inte har lika tydliga kopplingar till bestämda ord. Här kan det även påpekas att användarna har direkt tillgång till ljudfilerna där prosodiska drag kan kontrolleras.

Själva uppslagsordet markeras sedan med fetstil och vinröd färg. Utdrag (1) visar framställningen samt återger en svensk översättning. Exemplet används sedan igen i en skärmbild som diskuteras i följande avsnitt (figur 1).

(1) Exempel 1 för uppslagsordet *er það*

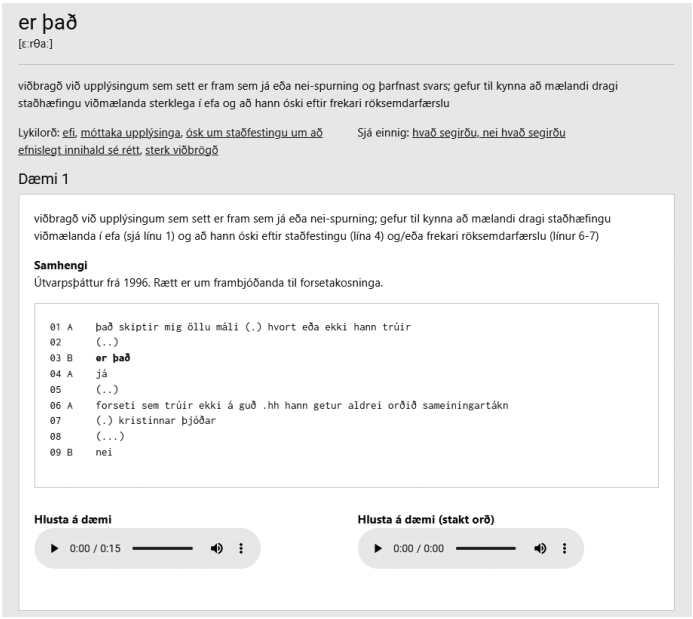
```
01 A  það skiptir mig öllu máli (.) hvort eða ekki hann trúir
      'det spelar mig en mycket stor roll   om han är troende eller inte'
02    (..)
03 B  er það
04 A  já
      'ja'
05    (..)
06 A  forseti sem trúir ekki á guð .hh hann getur aldrei orðið
      'president som inte tror på gud .hh han kan aldrig bli'
07    sameiningartákn kristinnar þjóðar
      'en samlande symbol för en kristen nation'
```

I utdraget förekommer två pauser som är mellan 0,4 till 1,9 sekunder långa, på rad 2 och 5. Vi delar in pauserna i tre typer, mycket korta (>0,3), vanliga pauser (ca 0,3–1,9) och långa pauser (ca <2,0). Pauserna mäts inte med en stor precision utan baseras snarare på vår upplevelse av längden jämfört med takten i samtalet, precis som samtalsdeltagarna gör själva. En allt för noggrann mätning av pauslängder skulle dra för mycket uppmärksamhet till siffror och försvåra användningen av transkriptioner för den allmänna ordboksanvändaren. På rad 6 förekommer sedan inandning, transkriberat *.hh*. I detta fall fungerar inandningen som en viktig övergångsmarkör mellan två led i ett argument. När vi transkriberar fler exempel kommer vi eventuellt att hitta fler fenomen som vi tycker behöver synas i transkriptionen. Tumregeln kommer dock alltid att vara att försöka göra en enkel och användarvänlig transkription.

Under insamlingen uppstod en del etiska frågor kring användningen av offentliga och privata samtal. Angående de privata samtalen har samtalsdeltagarna (och deras förmyndare) skrivit under informerat samtycke som tillåter att korta utdrag läggs ut på webben med ljudinspelningar. De offentliga samtalen ligger ute på webben och är fritt tillgängliga för alla. Redaktören kontaktade alla programledare per e-post och fick skriftligt tillstånd från dem att använda det material som finns på nätet. Vid valet av utdrag diskuterade arbetsgruppen även vikten av att inte välja utdrag som kan såra eller skada informanterna själva eller de som omtalas. Namn på ej offentliga personer är fingerade i transkriptionerna och de riktiga namnen rensades bort ur ljudfiler.

6. Webbsidans innehåll

Samtalsordboken bygger på en databas som knyter ihop uppslagsord med en rad olika komponenter: ljudinspelningar, transkribering, information om inspelningsår och annan relevant kontextuell information, en definition, nyckelord och länkar till uppslagsord som har liknande funktioner. Figur 1 visar uppslagsordet *er það* 'på riktigt' som det ser ut på samtalsordbokens webbsida.



FIGUR 1. Uppslagsordet *er það* i *Samtalsorðabók*.

Figur 1 visar det första samtalsutdraget av sju som finns under uppslagsordet *er það*. Överst framgår en fonetisk transkription av uppslagsordets uttal (IPA). Sedan kommer en rad nyckelord (isl. *lykilorð*) som kan förknippas med ett eller flera utdrag: TVIVEL, MOTTAGANDE AV INFORMATION, ÖNSKAN OM BEKRÄFTELSE, STARK RESPONS. Användaren hänvisas även till andra uppslagsord som har överlappande funktioner (*hvað segirðu/nei hvað segirðu* 'va'). Under nyckelorden visas uppslagsordets första exempel (isl. *dæmi*). Överst kommer en förklaring som i svensk översättning lyder: "mottagande av information som presenteras som en ja/nej-fråga; indikerar tvivel och visar att samtalspartern är av en annan

åsikt och önskar närmare förklaring". Efter förklaringen kommer information om inspelningen. Utdraget i fråga spelades in 1996 och det rör sig om ett radioprogram där man diskuterar presidentvalet på Island samma år. Samtalet skrivs ut som en dramadialog där alla ljud och pauser återges. Raderna numreras och samtalsdeltagarna får bokstäverna A och B. Längst ner finns sedan ljudfilerna, dels samtalsutdraget som en helhet (isl. *Hlusta á dæmi* 'lyssna på exempel'), dels själva uppslagsordet klippt ur sin kontext (isl. *Hlusta á dæmi (stakt orð)* 'lyssna på exempel (utan kontext')).

7. Avslutning

I denna artikel har jag presenterat projektet *Samtalsorðabók* som är en ny webbordbok som fokuserar på pragmatiska markörer i isländska samtal. Även om arbetet just har påbörjats finns webbordboken redan ute på Árni Magnússon-institutets hemsida. Tanken med att göra ordboken tillgänglig i ett tidigt skede är att lemmalistan och exemplen även i oredigerad form kan vara av stor nytta för språkinlärare. Arbetet kommer att fortsätta för öppen ridå. När vi för in nya ord och förklaringar eller lägger till fler exempel och nyckelord för redan existerande uppslagsord kommer det genast att synas på webbsidan.

För framtiden kan man också fundera på om tekniken kan ge oss nya möjligheter. En av de stora utmaningarna vi har funderat kring handlar om andraspråksinlärare och hur vi kan hjälpa dem att hitta de uppslagsord de söker. För en andraspråkstalare som stöter på en frekvent pragmatisk markör i samtal kan det vara svårt att veta hur den faktiskt stavas. Diskurspartiklar och andra frekventa markörer i samtal kännetecknas av fonetisk reduktion, t.ex. *þúst* eller *st* i stället för *þú veist* 'du vet'. Även för en modersmålstalare är stavningen inte alltid självklar, t.ex. när det gäller nya lån från engelska. Ska man t.ex. skriva in *ómægod*, *ómægad* eller *oh my god* i sökrutan? Ett av de förslag som har diskuterats är att utnyttja språkteknologiska resurser, t.ex. taligenkänningsprogram som har programmerats för att förstå isländska uppslagsord som uttalas med eller utan brytning. Man kan också tänka sig att andraspråkstalare vill kunna välja metaförklaringar på isländska eller engelska. De teknologiska möjligheterna är många och det kommer att bli spännande att se vart framtiden kan leda den nya isländska samtalsordboken.

Referenser

Andersen, Gisle 2014. Pragmatic borrowing. *Journal of pragmatics* 67, 17–33.

Bjarnason, Vilhjálmur 2020. ”Þú veist” hvurn fjandann? *Morgunblaðið* 30/4 2020.

Couper-Kuhlen, Elizabeth & Margaret Selting 2018. *Interactional Linguistics: study language in social interaction.* Cambridge: Cambridge University Press.

Hansen, Carsten 2015. Beskrivelsesproget i *Ordbog over Dansk Talesprog. LexicoNordica* 22, 57–76.

Hilmisdóttir, Helga 2001. Partiklarna *þúveist* och *skiluru:* ett isländskt ungdomssamtal under lupp. I: Nordenstam, Kerstin & Kerstin Norén (red), *Språk, kön och kultur. Rapport från fjärde nordiska konferensen om språk och kön i Göteborg den 6–7 oktober 2000*. Göteborg: Göteborgs universitet, 124–131.

Hilmisdóttir, Helga 2007. *A sequential analysis of* nú *and* núna *in Icelandic conversation.* (Nordica Helsingiensia 7.) Helsingfors: Helsingfors universitet.

Hilmisdóttir, Helga 2016. Responding to informings in Icelandic talk-in-interaction: A comparison of *nú* and *er það*. *Journal of Pragmatics* 104, 133–147.

Hilmisdóttir, Helga 2021a. Talspråkskorpusar, diskurspartiklar och lexikografi. *LexicoNordica* 28, 79–100.

Hilmisdóttir, Helga 2021b. Leikjatölvur og orðaforði unglinga. Rannsókn á framandorðum í orðaforða tveggja grunnskóladrengja. *Ritið* 3/2021, 117–144.

Íslensk orðabók 2002. Mörður Árnason (red.). Reykjavík: Edda.

Íslensk nútímamálsorðabók. Halldóra Jónsdóttir & Þórdís Úlfarsdóttir (red.). Árni Magnússon-institutet för isländska studier. <islenskordabok.is>. Hämtat augusti 2022.

Lindström, Jan 2008. *Tur och ordning. Introduktion till samtalsgrammatik*. Stockholm: Norstedts Akademiska Förlag.

ODT = *Ordbog over Dansk Talesprog*. Hansen, Carsten (red.). <odt.hum.ku.dk> Hämtat augusti 2022.

Opsahl, Toril 2015. Kan ord i bruk bli ord i bok? Urbane ungdomsvarieteter i framtidige ordbogsressurser. *LexicoNordica* 22, 131–149.

Samtalsorðabók. <samtalsordabok.arnastofnun.is>. Hämtat augusti 2022.

Steensig, Jakob 2004. *Sprog i virkeligheden. Bidrag til en interaktionel lingvistik*. Aarhus: Aarhus universitetsforlag.

Svavarsdóttir, Ásta 2007. Talmál og málheildir – talmál og orðabækur. *Orð og tunga* 9, 25–50.

When the users jump to conclusions. Presenting prescriptive information

Kristín Ingibjörg Hlynsdóttir & Kristín Bjarnadóttir

The topic of this paper is a method of presenting acceptability to the users of the online version of the Database of Modern Icelandic Inflection (DMII), with a short description of the classification used and a reference to a survey of one week of online queries, a total of 117,685 searches. The DMII was originally descriptive, and the inclusion of non-standard inflectional and spelling variants is known to confuse users who expect prescriptive data. Prescriptive information is provided in usage notes presented with the paradigms, but the users are apt to stop at the search list and jump to conclusions on acceptability without reading the notes. Non-standard headwords therefore need to be marked in the search list itself, with cross references to the standard forms, as needed.

KEYWORDS: morphology, inflection, Icelandic, language standard, language technology resource

1. Introduction

The DMII is an online reference for the general public and a resource for language technology (LT). The project has been ongoing at the Árni Magnússon Institute for Icelandic Studies (AMI) since 2002. The website (bin.arnastofnun.is) shows full paradigms of over 333,000 headwords and the data is available as downloadable CSV files. A smaller prescriptive version is available through an application programming interface (API), The DMII Core (Bjarnadóttir & Hlynsdóttir 2020). The DMII is an important resource for Icelandic LT and the website is very popular among the general public, with over 7 million page views in 2021.

The DMII was originally descriptive and the purpose was to show language use "as is", i.e., both standard and non-standard usage, with LT use in mind. The inclusion of non-standard inflectional and spelling variants can, at times, confuse the users of the website, as they expect a source from AMI to show only what is correct, i.e. prescriptive data. This was coun-

tered by adding usage notes with paradigms, mostly to guide users when choosing between inflectional variants.

In 2019, a new version of the DMII was released with extended usage analysis. A new grading and classification system made it possible to add more non-standard forms and to grade and differentiate between standard and non-standard forms, to create the DMII Core and for other LT uses. As a result, many more non-standard forms are now displayed on the website, both inflectional variants and headwords. This has led to more usage notes being added to the paradigms, which partly solves the problem of guiding the users as to good usage, i.e., in the choice of variant inflectional forms within the individual paradigm. The choice between headwords needs to be addressed in a different way, as non-standard headwords need to be marked in the search list, with a cross reference to the standard form. This is doubly important, as the users tend to forget the descriptive nature of the DMII and regard all the data therein as correct or acceptable. Users are also apt to stop at the search list and when they do so they never discover the usage notes in the paradigms, i.e., they jump to conclusions about acceptability.

The topic of this paper is a method of presenting acceptability to the users as efficiently as possible, with a short description of the classification used and a reference to a survey of one week of online queries, a total of 117,685 searches.

2. Descriptive vs. prescriptive

Over 20 years ago, the original purpose of the DMII was use in LT, at a very early stage of that field in Iceland. As Icelandic is a heavily inflected language, the immediate need was for data for search engines, etc., containing as large a vocabulary as possible with all corresponding inflectional forms. The first version of the DMII, published in 2004, contained an average of 27 inflectional forms per paradigm. Inclusiveness was also an important feature, which is why the DMII had to be descriptive and not prescriptive. The DMII data includes the good, the 'not-so-good' and the downright erroneous, according to the Icelandic language standard.

The first version of the data contained no classification of acceptability, and broadly speaking, its main function was to link lemmas and inflectional forms for use in LT. The first online version was a side product to

the LT data, but it has gained in importance and it is now used extensively by the public. The online users' need for prescriptive data is unquestionable, and the same applies to today's LT uses, such as spell checkers, grammar checkers, and any kind of language production, such as translation services, query systems, etc. The gradual change of the DMII from purely descriptive data to prescription with a reference to the Icelandic language standards is described in NSL 15 (Bjarnadóttir & Hlynsdóttir 2020). The development of the DMII is described in detail on the DMII website.

3. The original search results

Headwords in the DMII are presented on the web with full paradigms and usage notes, based on the classification described in NSL 15 (Bjarnadóttir & Hlynsdóttir 2020). Headwords can be searched for using a search bar on the website and if the search string returns multiple headwords, the results are listed as shown in figure 1, as headword, word class and domain.

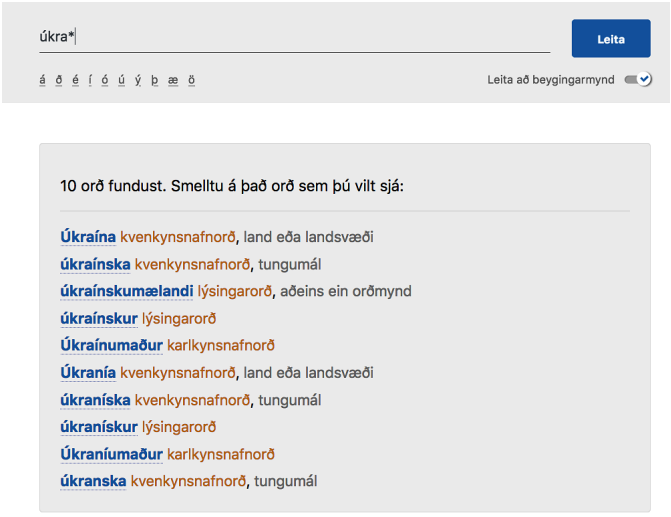
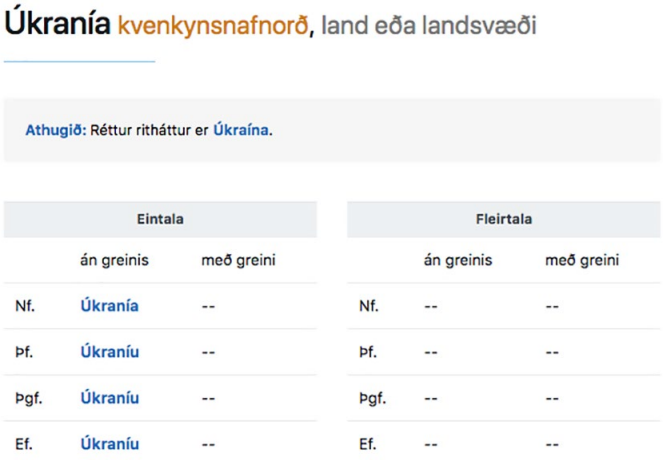


FIGURE 1. Searching for *úkra** in the online DMII.

Úkraína is the standard Icelandic form of the country name *Ukraine* but an alternative non-standard spelling variant is *Úkranía*. The headwords in figure 1 are all derived from these two variants, but the search result shows no indication of acceptability and the user might think that the variants *Úkraína* and *Úkranía* are equally acceptable, as they are both included in the

DMII. The user needs to click on a headword to see the paradigm in order to read the accompanying note to know whether the word form is correct or not. Selecting the word *Úkranía* from the list in figure 1 shows the paradigm, with a usage note saying that the correct spelling is *Úkraína*, cf. figure 2.



Úkranía kvenkynsnafnorð, land eða landsvæði

Athugið: Réttur ritháttur er **Úkraína**.

	Eintala		Fleirtala		
	án greinis	með greini		án greinis	með greini
Nf.	Úkranía	--	Nf.	--	--
Þf.	Úkraníu	--	Þf.	--	--
Þgf.	Úkraníu	--	Þgf.	--	--
Ef.	Úkraníu	--	Ef.	--	--

FIGURE 2. The paradigm of the word *Úkranía* with the usage note: "The correct spelling is ***Úkraína***."

Only selected parts of the classification for correctness produced for LT are displayed on the web as some of this data is specific to LT tasks and not suited for use by the general public. The classification is used as a tool to help create the usage notes which can be as long as needed as there is no need to save space. The notes can also contain references, i.e., links to further explanations. The aim is to make the notes as clear and readable as possible. The problem remains that many users only go as far as the search results, even though the explanations on the front page of the website clearly state the descriptive nature of the DMII.

The original source for headwords in the DMII was mostly lexicographical material, containing common non-standard word forms and spelling. With recent additions from the Gigaword Corpus (Steingrímsson et al. 2018) and work on error analysis, erroneous forms are deliberately added to the database with corrections to make sure to cover all common word forms for LT uses. This new type of data is of two types, i.e., errors in individual word forms (referred to as *non-standard word forms*)

and errors encompassing whole paradigms (referred to as *paradigms of errors*). These errors will be searchable with links to the correct forms online. It is important to make sure the users get a clear message that they are being referenced to a new word, so that they can see that the original form in their search string is not the standard form.

4. Learning from user search strings

Work on the DMII has been more LT focused in the last couple of years, with recent additions being produced as part of the Language Technology Programme for Icelandic (Almannarómur 2022). In order to make better use of the new data on the DMII website, the actual queries on the website were reviewed with the aim of finding out what the users were really searching for. All search queries from the week Jan. 24–30 2022 were collected, a total of 117,685 searches; 34,186 unique strings, with 12,021 strings not found in the DMII.

Many of the search strings not found in the DMII were ordinary, acceptable words missing from the DMII. As a result, approx. 2,500 new headwords were added to the database, using the Gigaword Corpus (Steingrímsson et al. 2018) for reference, and also adding some related compounds found in the Gigaword Corpus but not in the search list.

The remainder of the strings not found in the DMII contained various kinds of errors. A large portion of them were multiword search strings, such as:

- Noun phrases: *rauður hestur* 'red horse', *tveir ungir menn* 'two young men'
- Verbs with infinitive marker: *að gráta* 'to cry'
- Particle verbs: *ráðast á* 'attack'
- Prepositional phrases: *til upplýsingar* 'for information'
- Grammatical features included in the search string: *miðstig gamla* 'comparative old'
- Miscellaneous multiword strings: *ostur með sinnepi* 'cheese with mustard'

Other error strings contained wrong character sets, foreign queries, non-alphabetic characters and symbols, etc.

The remainder were real language errors, i.e., recognizable Icelandic word forms containing errors in spelling, word formation, typos, etc. These were analysed and classified according to the previously established system and then added to the DMII, as full-scale paradigms visible on the web or paradigms of errors and non-standard word forms. In the case of paradigms of errors and non-standard word forms, the corresponding correct headword and paradigm was sometimes missing from the DMII and had to be added.

The analysis of the search strings for words already in the DMII gave indications of the purpose of the search, which usually seemed to be for inflection, spelling or even word formation, although some of the strings are a bit harder to interpret. In the case of inflection, the users need to access the full paradigm, but in other cases the users may decide to make do with the search list, which means they will not see the needed notes on acceptability. Analysing the search strings gives limited scope for interpretation, and doing a thorough user survey would be very interesting. The simple analysis of the search strings described here does, however, confirm the need for a clearer presentation of the standard spelling in the DMII and the importance of including as many headwords as possible, including non-standard ones. The key issue is making it as easy as possible for the users to access the information.

5. Changes to the presentation of non-standard forms

As previously stated, many users seem to believe that everything found on the website is correct because they expect the data to be prescriptive. This is known from feedback given in e-mails, on social media, etc. The DMII describes the language "as is" and the scope of the DMII is much larger than any part of the Icelandic language standard. The DMII is, however, anything but exhaustive, either in vocabulary or inflectional forms. Some users have misconceptions about words not found in the DMII and they assume that words missing from the database have been deemed incorrect by the editors. The users have a tendency to regard the DMII as a language standard for Icelandic, which it is not. Some readers even expect the DMII to be exhaustive and assume missing words to be nonexistent in the language, which is certainly not the case.

The DMII editorial concept has been that the users must be able to find non-standard words and inflectional forms and the aim is also to explain

why they are not considered acceptable, as far as possible. The vocabulary of the DMII (or any other source) is not exhaustive and it never will be, but the goal is to include as much as possible.

5.1. Search heads

After making sure the users find what they are looking for, the next task is ensuring that they actually understand the given information, preventing them from jumping to conclusions. As shown in figure 1 in section 3, search lists were misleading for users that did not proceed to the usage notes. Adding data in a shorter form (search heads) to the search lists themselves solves this problem. The search heads are standardised based on style/register and grade but semi-manually added to each word. The headers also contain word class and domains.

Examples of search heads are shown in the following tables:

TABLE 1. Corrections, (usually) showing target word or word form.

Úkranía kvenkynsnafnorð. Réttur ritháttur er **Úkraína**.	[Correct form]
Egilssaga kvenkynsnafnorð. Hefðbundinn ritháttur er **Egils saga**.	[Traditional form]
Hercules karlkynsnafnorð. Íslenski ritátturinn er **Herkúles**.	[Icelandic form]
ábrestur kvenkynsnafnorð. Afbrigði af **ábrystir**.	[Variant form]
kólumbín hvorugkynsnafnorð. Eldra heiti á **níóbín**.	[Older term]
kjurr lýsingarorð. Framburðarmynd af **kyrr**.	[Pronunciation form]
hör hvorugkyn. Rétt er að hafa orðið í karlkyni.	[The correct gender is masculine]

TABLE 2. Words that are not fully acceptable, without direct reference to a standard form.

sinnhver óákveðið fornafn. Ekki viðurkennt mál.	[Unacceptable]
selebreita sagnorð. Sletta.	[Unacceptable loanword]
Discorites karlkynsnafnorð. Erlendur ritháttur.	[Foreign spelling]

The search heads are also used for references between equally correct forms or sets of easily confused words, such as homophones.

TABLE 3. Homophones.

skrýtinn lýsingarorð. Einnig ritað **skrítinn**.	[Also written]
híði hvorugkynsnafnorð. Orðið **hýði** hefur aðra merkingu.	[Confusion set]

Style or register is simplified for presentation in the search list, using "Gamalt" 'old' when the style is obsolete, Old Icelandic or old-fashioned, and "Sjaldséð" 'rare' for dialectal, poetic or rare words. The actual usage notes with the paradigms usually contain more specified data, and these are individually written for each paradigm.

TABLE 4. Age, style or register.

afbatan kvenkynsnafnorð. Gamalt	[Old (classified obsolete)]
mánadagur karlkynsnafnorð. Gamalt	[Old (classified Old Icelandic)]
bedraga bedró bedrógum bedregið sagnorð. Gamalt	[Old (classified old-fashioned)]
aflaslór hvorugkynsnafnorð. Sjaldséð	[Rare (classified dialectal]
aldurlok hvorugkynsnafnorð. Sjaldséð	[Rare (classified poetic)]
afarvogun hvorugkynsnafnorð. Sjaldséð	[Rare (classified rare)]

The new version of the search results for *úkra** are shown in figure 3, marking the acceptability of the spelling variants with "Réttur ritháttur er ... " 'The correct spelling is ... ', cf. figure 1 in section 3 where the search heads are not shown. This immediately shows the users which spelling variants are standard and which are not.

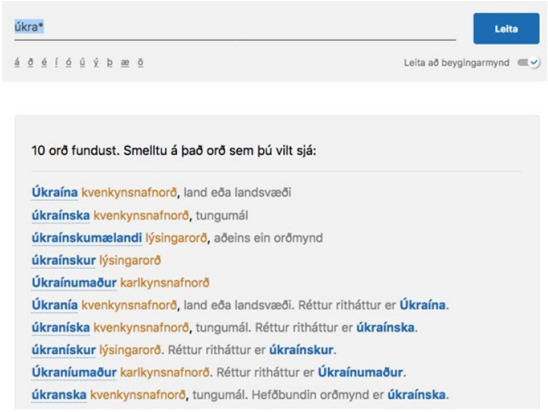


FIGURE 3. Searching for *úkra**, showing search heads.

5.2. Non-standard word forms and paradigms of errors

Neither the non-standard word forms nor the paradigms of errors are visible on the DMII website but they are to be used as supplementary data in the search to the benefit of the users. The goal is to make users aware of what is standard language and what is not, and the results from the non-standard word forms and paradigms of errors are only presented as references to the correct form. As of September 2022, this is still not visible for web users but will be added soon. The results will appear with the regular DMII results with the header "Þú gætir átt við ... " 'You might be looking for ... '.

Both headwords and inflectional forms are searchable in the DMII and search for inflectional forms returns a list of headwords containing that form. If a search string is found in both a regular DMII headword and as a non-standard word form or in a paradigm of errors, the results for the non-standard form are shown beneath the regular results. If the string is only found in the non-standard data, the results are beneath the standard message saying it was not found in the database. As an example, the form "alnar" is an inflectional form of the feminine noun *öln* 'ulna' but it is also a possible error form of four other words. In this case, the word *öln* would be listed as usual but after that, the other four headwords are listed as possibilities, with the caption 'You might be looking for':

öln kvenkynsnafnorð

Þú gætir átt við 'You might be looking for':
ala ól ólum alið, sagnorð
alin kvenkynsnafnorð
alinn lýsingarorð
álna álnaði álnað, sagnorð

The aim is to show the users that their search string is a standard form of the first word but only similar to (or an error form of) the other four. It is then up to the user to determine which headword they are actually looking for.

5.3. Unsolved problems

Questions of word boundaries are the reason for some of the most common types of real language errors in the search strings. These are difficult

to cope with in the DMII, which was originally strictly based on single-word paradigms.

The first type of error is splitting compounds in the search strings. This type of error is outside the scope of the DMII at present since all possible variations of erroneously split compounds cannot be added to the data. Using a compound splitter (Daðason et al. 2020) in "reverse mode" on the search strings might work, in the form of a suggestion, similar to how the non-standard word forms and paradigms of errors are presented. Examples of split compounds in the search strings are *lyfja afhending* for *lyfja-afhending* 'delivery of drugs', and *almennings stöðum* (dative), for *almenningsstöðum* 'public places'.

The other type is joining words erroneously, as in writing prepositional phrases and phrasal adverbials as continuous strings: *afhverju* (prepositional phrase, incorrect) for *af hverju* (correct form) 'why'; *einskonar* (adv., incorrect) for *eins konar* (correct form) 'some kind of'. Errors are also common in word class dependant word boundaries, as specified in the Spelling Rules: *ofstór* adj. (incorrect) for *of stór* (correct form) 'too big'. (The spelling rule specifies that the adverb *of* 'too' is concatenated to nouns and verbs, but a free form preceding adjectives and adverbs.)

The problem is that multiword headwords cannot be added in all possible cases, but work is in progress on finding a method of presenting suggestions based on current LT tools for Icelandic in order to give the users hints on the nature of their search string errors, in the form: "You might be looking for ..." or "The correct form might be ...".

6. Conclusion

This paper has described some of the steps taken recently in the development of the DMII from being purely descriptive towards being prescriptive. In the last few years, the focus of the project has been on LT, but it is now shifting to the users of the website, looking at what they really search for and how they may be misreading the information. The remedy proposed in this paper is to shift as much information as possible to the earliest possible place in the search process. In that manner, the search heads and added references from errors to standard forms in the search lists should help users looking to correct their grammar and spelling, even when they might have a tendency to jump to conclusions.

References

Almannarómur. <https://almannaromur.is/en>. Accessed September 2022.

Beygingarlýsing íslensks nútímamáls. [*The Database of Modern Icelandic Inflection*.] <bin.arnastofnun.is>. (no date) Kristín Bjarnadóttir, editor. The Árni Magnússon Institute for Icelandic Studies.

Bjarnadóttir, Kristín, & Hlynsdóttir, Kristín Ingibjörg. 2020. Online Data on Icelandic Inflection: Descriptive to Prescriptive: "Why, for whom, by whom" and how? *Nordiska studier i lexikografi* 15. Rapport från 15 konferensen om lexikografi i Norden. Helsingfors 4–7 juni 2019, pp. 71–79.

Daðason, Jón Friðrik, Mollberg, David Erik, Loftsson, Hrafn, Bjarnadóttir, Kristín. 2020. Kvistur 2.0: a BiLSTM Compound Splitter for Icelandic. *LREC 2020 Proceedings,* pp. 3984–3988.

Steingrímsson, Steinþór, Helgadóttir, Sigrún, Rögnvaldsson, Eiríkur, Barkarson, Starkaður, & Guðnason, Jón. 2018. Risamálheild: A Very Large Icelandic Text Corpus. *Proceedings of LREC 2018*, pp. 4361–4366. Myazaki, Japan.

Okynniga pluraler. Normering och bruk av s-plural speglat i SAOL och SO

Louise Holmer & Kristian Blensenius

In this article, we discuss a noun declension in Swedish indicating plural with the suffix *-s*. The suffix has been debated among language users, grammarians, language planners and lexicographers on and off for several decades. The *-s* suffix is, in some variants, considered to belong to colloquial speech, for example in forms like *avokados, bikinis* and *cellos* ('avocados', 'bikinis', 'cellos/celli', respectively), whereas formal written Swedish is said to prefer plural forms such as *avokado-r/-er*, *bikini-er* and *cello-r*. However, there seems to be consensus that s-plurals are useful, and these forms are now more or less officially accepted.

We examine the use of s-plural for some widely debated loan words in modern Swedish newspaper texts, and relate them to grammatical statements, recommendations from language planners and recommendations in two Swedish monolingual dictionaries, namely *The Swedish Academy Glossary* (SAOL, 14th ed. 2015) and *The Contemporary Dictionary of the Swedish Academy* (SO, 2nd ed. 2021). Areas for improvement in the dictionaries are identified, and it is suggested that inflectional suppletion is considered for some s-plural definite forms.

NYCKELORD: morfologi, s-plural, suffix, svenska ordböcker

1. Inledning, bakgrund och syfte

När *Svensk ordbok utgiven av Svenska Akademien* publicerades i en reviderad upplaga i maj 2021 (SO 2021), var en av nyheterna att fler ord än tidigare var försedda med pluralböjning med *-s*, t.ex. *avokado*, med pluralangivelsen "*avokados* eller *avokador*". Även ett nytillagt ord som *hashtag* hade fått s-plural som förstaform (*hashtags* eller *hashtaggar*). Pluralformen *avokados* är numera vanlig i tal- och skriftspråk, medan *avokador* är den (enda) form som rekommenderas i t.ex. *Svenska Akademiens ordlista* (SAOL 14, 2015). Förutom nämnda former används också pluralen *avokadosar*, även om den framför allt återfinns i talspråket och det lediga skriftspråket. Den senare pluralböjningen redovisas i SAG (2:83, 104) och kallas ibland *-sar*-plural (Josefsson 2018), och den känns igen i

former som *bikinisar*, *containersar* och *kängurusar*. Språkvårdare avråder vanligen från *-sar*-plural i vårdat skriftspråk.

I den här artikeln ger vi exempel på lånord där det svenska skriftspråket uppvisar variation i fråga om pluralböjningen, med tonvikt på s-plural (’okynniga’ eller ’oregerliga’ pluralformer.) Vi undersöker hur dessa ords pluralformer behandlas i de två svenska enspråkiga ordböckerna SAOL 14 och SO 2021, och dessa undersökningar kontrasteras mot normer och rekommendationer som uttrycks av framför allt Språkrådet, liksom mot grammatik i bl.a. *Svenska Akademiens grammatik* (SAG). Syftet är att jämföra språkvårdsrekommendationer med språkbruket vid ett urval av ord där pluralformerna sedan tidigare har konstaterats variera. Undersökningarna bidrar till att belysa ett lexikografiskt specialområde – angivelse av numerusböjning – med koppling till såväl grammatiska regler och språkvårdens normering som till språkbruket. Resultaten har relevans för det praktiska lexikografiska arbete som vidareutvecklingen av SAOL och SO innebär, liksom för andra ordböcker och lexikografiskt arbete i stort.

S-plural är inte något nytt fenomen i svenskan, utan har använts sedan 1700-talet (och i vissa fall tidigare, jfr Söderberg 1983). Trots det diskuteras s-plural återkommande i språkvårdssammanhang även i våra dagar och är också föremål för lexikografisk uppmärksamhet vid revidering av ordböcker. I Söderberg (1983) beskrivs ett antal undersökningar av s-plural i svenskan mot bakgrund av attityder till ”svengelska” hos svenska språkvårdare, liksom mot bruket av s-plural i svensk text. Dessutom görs i Söderberg (1983) bland annat en genomgång av olika ordböckers hantering av dessa former. Sammanlagt sexton ordböcker undersöks med avseende på deras respektive (eventuella) registrering av s-plural i fråga om totalt 295 ord (Söderberg 1983:198–205). Av Söderbergs genomgång kan nämnas ett mycket litet urval ord som *bolero*, *inka*, *sandwich* och *scone*, som i någon av ordböckerna vid något tillfälle har försetts med pluralform på s.[1]

I denna artikel går vi igenom ett urval fall av s-plural i SAOL och SO och jämför ordböckernas lösningar med språkvårdens rekommendationer. Först presenteras de undersökta orden och ordböckerna. Relevant grammatisk och språkvårdande litteratur presenteras också helt kort (avsnitt 2). Därefter redogörs närmare för vad den grammatiska och språkvår-

1 I SAOL 14 (2015) och SO (2021) är uppslagsformen *scones*, dvs. plural, i stället för som tidigare *scone*.

dande litteraturen säger om s-pluralen (avsnitt 3) och vad som anges i några SAOL-upplagor och SO-upplagor. Slutligen presenteras ett förslag på böjning vid ord som kan kopplas till mer än ett paradigm, grundat i s.k. saxade paradigm, och artikeln sammanfattas.

Artikelförfattarna är verksamma vid Institutionen för svenska, flerspråkighet och språkteknologi, Göteborgs universitet, där ordböckerna SAOL och SO utarbetas sedan många år. Undersökningarna utgör ett av flera områden där skillnader mellan de båda ordböckerna inventeras i syfte att minska andelen omotiverade sådana i samband med revideringsarbetet.

2. Material för undersökningarna

De undersökta orden är valda utifrån att de ofta är föremål för språkvårdens insatser i fråga om böjningsmönster, liksom för att de ofta utgör intressanta variationsexempel för lexikografer. I artikeln fokuserar vi på de tre orden *avokado*, *bikini* och *hashtag*. Orden *avokado* och *bikini* är välkända i svenskan sedan åtminstone 1930- och 1940-talet, medan *hashtag* utgör exempel på ett av svenskans nyare ord, med förstabelägg i SO från år 2010.

Förutom dessa tre ord är förstås den sammantagna mängden ord med möjlig pluralböjning med *-s* mer omfattande. Övriga exempelord som nämns i artikeln har hämtats från Språkrådets Frågelåda (som finns allmänt tillgänglig på nätet) och från översikter i Josefsson (2018). Urvalet av ord är menat att belysa olika typer av böjningsvariation som medför olika typer av morfologisk-lexikografiska överväganden för lexikografer. För generella genomgångar av s-plural och mer omfattande exempelsamlingar hänvisas till Söderberg (1983) och Josefsson (2018).

Det material som ligger till grund för undersökningarna utgörs av svenska, enspråkiga ordböcker i form av SAOL 14 (2015) och SO (2021). Den referensgrammatik som används är *Svenska Akademiens grammatik* (SAG, 1999), men vi jämför även mycket kort med senare grammatikor som Hultman (2003), Josefsson (2009), Bolander (2012) och Lundin (2014). De språkvårdande skrifter som används utgörs av det svenska Språkrådets rekommendationer (Frågelådan) liksom *Svenska skrivregler* (Karlsson 2017) och *Språkriktighetsboken* (2016).

Några viktiga skillnader mellan de båda ordböckerna är att SAOL 14 är mer normativ och SO 2021 mer deskriptiv (Blensenius, Holmer & Sköldberg 2021:41). SAOL har också en längre utgivningstradition än

SO; SAOL:s första upplaga kom ut år 1874 som en sorts ram till den historiska ordboken *Svenska Akademiens ordbok* (SAOB). Därefter har ordlistan publicerats relativt regelbundet i tryckt form med den fjortonde och senaste upplagan år 2015. För en bakgrund till SAOL i allmänhet hänvisas till Gellerstam (red.) 2009, där många olika ingångar till tidigare upplagor av ordlistan ges. SAOL 14 beskrivs i t.ex. Malmgren (2014).

SO 2021 är en vidareutveckling av den tryckta ordboken SO 2009, som i sin tur bygger på *Nationalencyklopedins ordbok* (NEO, 1995–1996) och den korpusbaserade ordboken *Svensk ordbok* 1986. SO 2021 publicerades enbart på nätet i ordboksportalen svenska.se och som app för iOS och Android, och alltså *inte* i form av en tryckt bok. I SO 2021 har det deskriptiva perspektivet renodlats jämfört med föregångaren. För en allmän genomgång av SO 2021, se Sköldberg (2022).

3. Pluralbeskrivningar i grammatiken och språkvården

I syfte att kartlägga olika infallsvinklar på böjning av inlånade substantiv ges i detta avsnitt en översikt över hur sådana ords pluralformer behandlas i referensgrammatiken SAG och några andra svenska grammatiska läroböcker samt i Språkrådets rekommendationer.

3.1. Grammatikornas beskrivningar

I svenskan har man traditionellt räknat med fem eller sex deklinationer för substantiven, baserat på substantivens böjning i plural. I SAG tillkom dock en sjunde deklination (SAG 2:63). Denna innehåller substantiv med suffixet *-s* i plural, t.ex. *dissenters* och *tricks*. Att den mest omfattande referensgrammatiken för svenska inkluderat deklinationen har, trots grammatikens deskriptiva inriktning (SAG 1:20), kommit att skänka viss legitimitet åt användandet av s-plural för ord som *avokado*, *bikini* och *hashtag*: *avokados*, *bikinis* respektive *hashtags*.

Sedan SAG gavs ut har vissa grammatikläroböcker för högskolenivå anammat s-pluraldeklinationen (se t.ex. Bolander 2012:114), medan andra inte nämner den (t.ex. Lundin 2014). *Svenska Akademiens språklära* (Hultman 2003:64–65) antar sex deklinationer. S-pluralen nämns förvisso, men någon särskild deklination för substantiv med denna böjning antas inte. Josefsson (2009) diskuterar s-pluralen som en möjlig

sjätte deklination (Josefsson räknar annars med fem deklinationer), men uttrycker samtidigt tveksamhet:

> Att ge ord med plural-*s* status som en egen deklination är tveksamt; man ska komma ihåg att många importord behåller sitt plural-s under en kortare övergångstid, och när ordet blir mer hemtamt i språket övergår det till att pluralböjas enligt någon av de andra deklinationerna. Kanske tycker många att pluralformer som *avokador* och *kiwier* ser konstiga ut i början, men ögat vänjer sig, och efter ett tag reagerar vi inte längre. (Josefsson 2009:70)

3.2. Språkrådets rekommendationer

De pluralformer som rekommenderas av Språkrådet i fråga om substantiv som *bikini*, *container* och *känguru* utgörs sällan av *-sar*, utan varierar mellan andra ändelser. Ibland tar orden nolländelse (*en bikini, flera bikini*), ibland *-rar* (*en container, flera containrar*) och i vissa fall *-er* (*en känguru, flera känguruer*) (jfr Josefsson 2018).

Språkriktighetsboken (2016:145–176) behandlar vissa frågor om pluralformers morfologi, t.ex. den om latinsk pluralböjning med *-a* (*centra*, *fakta* m.fl.), men ämnet s-plural berörs inte. I den senaste upplagan av *Svenska skrivregler* (Karlsson 2017) har plural med *-s* givits en framträdande plats i ett eget avsnitt (7.1.1), något som inte var fallet i den föregående upplagan, *Svenska skrivregler* från 2008, där s-pluralen i en underavdelning till underavsnitt 7.2.5, ”Böjning av vissa främmande substantiv”, helt kort nämns och avfärdas som olämplig p.g.a. att den inte kan sättas i bestämd form.

Vid en jämförelse mellan SAG (se 3.1 ovan) och Språkrådets *Svenska skrivregler* (Karlsson 2017) blir det också tydligt att *Svenska skrivregler* går ett eller ett par steg längre än SAG i fråga om pluralformens etableringsgrad. Medan SAG (2:79) återhållsamt anger att ”Bruket av *-s* har i de flesta fall en relativt osvensk prägel”, menar *Svenska skrivregler* (Karlsson 2017:104) att ”S-plural är ganska vanligt förekommande i svenskan” och ”För vissa ord är […] s-pluralen så etablerad att den åtminstone i vissa sammanhang dominerar helt över andra böjnings-

mönster". Språkrådet går i Frågelådan[2] ännu ett steg i fråga om s-pluralen:

> […] eftersom den har en lång historia i svenskan och eftersom vissa svenska ord (*sambos*, och lite skämtsamt *snyggos*, *gubbs*, *kvinns*) och lånord från språk med andra pluralformer ibland får s-plural (*martinis*, *igloos*, *samosas*, *hijabs*), ser vi det nu som rimligt att betrakta den som en del av det svenska språksystemet.

Ett vanligt argument emot s-plural i dag[3] återfinns i *Svenska skrivregler*: "S-pluralmönstret saknar en etablerad form för bestämd form plural" (Karlsson 2017:104). Tanken verkar vara att s-plural-deklinationen ska hållas intakt; *s*-suffixet i obestämd form plural ska följa med även i bestämd form plural. Strängt taget bör därmed ett substantiv i obestämd form plural som *bikinis* skrivas *bikinisarna* eller möjligtvis *bikinisen* i bestämd form. Språkrådet anger emellertid i Frågelådan[4] att plural med *-sar* (som i *bikinisar*) i skriftspråk kan uppfattas som talspråkligt och rekommenderas inte i det vårdade skriftspråket.

En s-pluralform i skrift utgör i normalfallet sällan något problem (*containers*/*containrar*). Det som i stället kan utgöra problem för skribenter är alltså framför allt när bestämd form plural aktualiseras (*containersen*? *containersarna*?). I sådana fall brukar språkvårdare rekommendera pluralsuffix som ansluter sig till det svenska böjningssystemet (*två containrar – de containrarna*).

4. S-plural i SAOL och SO jämfört med bruket i text

I svenska ordböcker anges substantivens numerusböjning vanligen i form av bestämd form singular och obestämd form plural (Svensén 2004:172), något som också görs i SAOL 14 och SO (2021). Tanken att användaren ska kunna sluta sig till övriga former från de angivna böjningsformerna.

---

2 "Hur ser Språkrådet på s-plural?" <https://frageladan.isof.se/visasvar.py?sok=plural-s&svar=79712&log_id=977183%3E>. Hämtat september 2022.

3 Tidigare har det även betraktats som problematiskt att s-pluralformen kommit att uppfattas som singularform, t.ex. *en jumpers*, vilket kunnat leda till pluraler som *jumpersar* (se t.ex. Wellander 1970:161).

4 "Hur ser Språkrådet på s-plural?" <https://frageladan.isof.se/visasvar.py?sok=plural-s&svar=79712&log_id=977183%3E>. Hämtat september 2022.

Enligt Svensén (2004) behöver böjningsvarianter framför allt redovisas i receptionsordböcker, medan produktionsordböcker rekommenderas att begränsa sig till den variant som är vanligast. Dock kan normativa ordböcker även behöva redovisa former som inte är helt accepterade för att upplysa om vilken form som är att rekommendera (Svensén 2004:165). I det följande ges en översikt över hur pluralformerna anges i den mer normativa SAOL respektive den mer deskriptiva SO.

4.1. Översikt över pluralvariation i några olika upplagor av SAOL

De aktuella orden *avokado*, *bikini* och *hashtag* har ibland försetts med olika rekommendationer avseende böjningssätt och böjningsformer över tid (jfr Josefsson 2009). Nedan följer några exempel på ord som har registrerats med olika pluralformer i olika upplagor av SAOL, dels *avokado* och *bikini* (*hashtag* är som nämnts nytillagt, så det redovisas inte här), dels *film* och *schlager*, eftersom dessa ord har genomgått flera (oväntade) byten av pluralformer. Se tabell 1.

TABELL 1. Pluralformer i ett urval av SAOL-upplagorna.

	SAOL 8 (1923)	**SAOL 9 (1950)**	**SAOL 10 (1973)**	**SAOL 11 (1986)**	**SAOL 13 (2006)**	**SAOL 14 (2015)**
avokado			*avokador*	*avokador* el. *avokadoer*	*avokador* el. *avokadoer*	*avokador*
bikini			*bikini*	*bikini* äv. *bikinier*	*bikini* äv. *bikinier*	*bikinier* hellre än *bikinis*
film	*filmer* el. *films*	*filmer* äv. *films*	*filmer*	*filmer*	*filmer*	*filmer*
schlager		*schlager* el. *schlagrar*	*schlager* el. *schlagrar*	*schlagrar*	*schlagrar* el. *schlager*	*schlagrar* el. *schlager* hellre än *schlagers*

När ordet *avokado* togs med som uppslagsord i SAOL 10 (1973) noterades pluralböjningen *avokador*. I upplaga 11–13 ges även formen *avoka-*

doer, medan *avokados* ännu inte har tagits med som pluralform i någon upplaga av SAOL.

Uppslagsordet *bikini* togs med i SAOL 10 (1973) med pluralböjningen *bikini,* alltså samma form som singular, och i SAOL 11 (1986) fanns också variantböjningen *bikinier* med. I SAOL 14 (2015) har pluralformen *bikini* försvunnit och ersatts av kommentaren ”*bikinier* hellre än *bikinis*”.

Ordet *film* togs med i SAOL 8 (1923) med pluralangivelsen ”*filmer* eller *films*”. I fråga om s-plural ser vi alltså här ett relativt tidigt exempel. Pluralformen på *-s* finns med även i SAOL 9 (1950) men tas bort i och med upplaga 10 (1973). I modern svenska utgör pluralformen en ickefråga med *filmer* och *filmerna* som allenarådande former i obestämd och bestämd form.

Schlager togs med som uppslagsord i SAOL 9 (1950) med pluralformerna ”*schlager* eller *schlagrar*”. I SAOL 13 (2006) har dessa pluralformers inbördes ordning bytt plats. I SAOL 14 (2015) finns för första gången s-formen med, uttryckt i rekommendationen ”*schlagrar* eller *schlager* hellre än *schlagers*”.

Denna korta genomgång av några exempelord och deras pluralformer i SAOL visar att den norm som uttrycks i ordböckerna kan ändras över tid (se också tabell 1).

4.2. *Avokado, bikini* och *hashtag* i de senaste upplagorna av SAOL och SO

De olika inriktningarna hos SAOL och SO visar sig bl.a. i att SAOL 14 har fler och mer explicita rekommendationer än SO 2021, t.ex. i form av kommentaren ”Använd hellre” vid somliga ord och böjningsformer (t.ex. ”**attachment** – Använd hellre *bilaga*”), liksom ”X hellre än Y” vid vissa sidoformer (t.ex. ”**skrolla** hellre än **scrolla**”) och böjningsformer (t.ex. vid **mysa**: ”*myste* hellre än *mös*”). Den mer deskriptivt inriktade SO 2021 tillåter t.ex. s-plural för *avokado*, medan den mer normativt inriktade SAOL 14 inte anger *avokados* som rekommenderad pluralform. Det kan noteras att ingen av ordböckerna föreslår eller på annat sätt nämner *-sar*-plural som möjlig form.

Pluralformer vid *avokado*, *bikini* och *hashtag* ser ut enligt följande i de båda ordböckerna:

SAOL: *avokador*
bikinier hellre än *bikinis*
hashtaggar hellre än *hashtags*

SO: *avokados* eller *avokador*
bikinis
hashtags eller *hashtaggar*

I uppställningen ovan kan noteras att SAOL alltså ger företräde för vissa pluralformer framför andra (markerat med ”hellre än”), medan SO jämställer de olika böjningsformerna.

I skriftspråksbruket varierar pluralböjningen av *avokado* mer än i ordböckerna, och följande pluralformer går att hitta i olika typer av texter i en enkel korpusundersökning[5]: *avokado*, *avokados*, *avokador*, *avokadoer*, *avokadon* och *avokadosar*. Alla dessa former är naturligtvis inte registrerade i ordböcker, och alla formerna är heller inte rekommenderade av svensk, samtida språkvård. I fråga om de uppräknade pluralformerna är *avokador* den mest högfrekventa i tidningstext, ungefär 10 gånger vanligare än *avokados*. (Stavningen varierar också mellan *avokado* och *avocado,* något som dock spelar mindre roll i sammanhanget.) I fråga om *bikini* är pluralvariationen i bruket inte lika stor, men det är ingen tvekan om att pluralformen *bikinis* är betydligt vanligare än den av SAOL 14 rekommenderade *bikinier*. I fråga om *hashtag* utgör *hashtags* den vanligare pluralformen i obestämd form, medan bestämd form har klar övervikt för *hashtaggarna* jämfört med *hasthtagsen* i tidningstext. Detta talar för att båda pluralformerna bör finnas med i ordböckerna. Språkbrukarna kan också lösa svårigheten med bestämd form plural genom att använda *hashtags* i obestämd form och *hashtaggarna* i bestämd form (se nedan).

Poängen är att båda ordböckerna behöver förhålla sig till språkbrukarna och de pluralformer som används framför allt i skrift. Samtidigt finns det en intressant skillnad mellan den mer normativa SAOL och den mer deskriptiva SO, och det kan finnas goda skäl att ge olika förstaformer i plural i de båda verken (jfr Svensén 2004). Dessutom behöver lexikograferna betänka

5 Den mycket enkla metoden har här utgjorts av sökningar på ordformer i nyhetstexter och texter från sociala medier i korpussökverktyget Korp (jfr Borin et al. 2012).

eventuella svårigheter för en ordboksanvändare som går till svenska.se och ser olika böjningsformer bredvid varandra vid samma uppslagsord.

5. Saxad böjning

Med stöd i fördelningen av pluralvarianter (se avsnittet ovan) vill vi diskutera en kompromisslösning: vi föreslår att det ska vara möjligt att presentera s.k. *saxad* böjning av den aktuella pluraltypen i ordböckerna (SAG 2:544), dvs. böjning där böjningsformerna för ett och samma substantiv kan föras till olika deklinationer. I SAG behandlas saxad böjning framför allt i fråga om verb som kan föras till olika konjugationer, t.ex. verbet *dyka* som kan böjas *dyka – dök – dykt* (jfr *dyka – dykte – dykt*). Inget hindrar alltså att språkbrukaren rör sig mellan olika paradigm i obestämd och bestämd form i text, t.ex. som i följande uppställning:

singular	plural obest.	plural best.
avokado	*avokados*	*avokadorna*
bikini	*bikinis*	*bikinierna*
hashtag	*hashtags*	*hashtaggarna*

Ordböckernas behandling av ord med möjliga saxade paradigm behöver inte frångå befintliga uppgifter om plural. Ingenstans föreskriver ordböckerna att den språkbrukare som använder pluralformerna *avokados*, *bikinis* och *hashtags*, dvs. s-former, i obestämd form också förbinder sig att låta s-formen ingå i plural bestämd form och skriva *avokadosen*, *bikinisen* och *hashtagsen*.

I skrift kan en läsare möjligen uppfatta skribenten inkonsekvent, ifall skribenten rör sig mellan paradigmen i en text, men troligtvis rör det sig om så pass få förekomster inom en och samma text att en läsare inte störs nämnvärt, om den ens noterar tilltaget.

6. Summering

Ord med pluralvarianter kan innebära valmöjligheter och ibland vålla problem för skribenter. Till vilken instans ska egentligen en skribent i behov av råd vända sig till? De okynniga pluralerna behandlas, som visats i artikeln, på delvis olika sätt i grammatikor, av språkvårdare, av språk-

brukare och i ordböcker. Trots allt är det språket i bruk som ligger till grund för både ordböckerna och grammatikböckerna, liksom språkvårdens olika rekommendationer.

För den praktiska lexikografin utgör variationen i pluralböjning ett intressant problem eftersom ordböcker vanligen baseras på *skriftspråks*bruket, både i fråga om urvalet av uppslagsord liksom hur uppslagsordens böjningsformer registreras, och i det senare fallet inte sällan med utgångspunkt i språkvårdens rekommendationer. I somliga fall, beroende på ordbokens inriktning, förses kanske också vissa böjningsformer med någon typ av normerande kommentar. De lexikografiska rekommendationerna i SAOL och SO behöver i vilket fall förhållas till språkvårdens – numera frikostigare – rekommendationer sett till s-plural. En mer normativ ordbok som SAOL har av tradition varit mer restriktiv sett till s-plural, medan SO (2021) har tagit med fler fall av sådana. I den fortsatta revideringen av ordböckerna behöver redaktionen t.ex. ta ställning till ifall inte också SAOL behöver anamma en mer tillåtande attityd till s-plural.

Referenser

Blensenius, Kristian, Louise Holmer & Emma Sköldberg 2021. SAOL 14 som rättesnöre – diskussion kring den senaste upplagan. *Lexico-Nordica* 28, 39–58.

Bolander, Maria 2012. *Funktionell svensk grammatik*. 3 uppl. Stockholm: Liber.

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

Gellerstam, Martin (red.) 2009. *SAOL och tidens flykt. Några nedslag i ordlistans historia*. Stockholm: Norstedts.

Hultman, Tor G. 2003. *Svenska Akademiens språklära*. Stockholm: Svenska Akademien.

Josefsson, Gunlög 2009. *Svensk universitetsgrammatik för nybörjare*. 2 uppl. Lund: Studentlitteratur.

Josefsson, Gunlög 2018. *Avokadosar* och *kepsar* – ett epentetiskt *s* med olika funktioner. *Språk och stil* NF 28, 5–21.

Karlsson, Ola (red.) 2017. *Svenska skrivregler*. 4 uppl. Stockholm: Språkrådet & Liber.

Lundin, Katarina 2014. *Tala om språk. Grammatik för lärarstuderande*. 2 uppl. Lund: Studentlitteratur.

Malmgren, Sven-Göran 2014. Svenska Akademiens ordlista genom 140 år: mot fjortonde upplagan. *LexicoNordica* 21, 81–98.

SAG 1–4 = Teleman, Ulf, Erik Andersson & Staffan Hellberg 1999. *Svenska Akademiens grammatik*. Del 1–4. Stockholm: Norstedts.

SAOB = *Ordbok över svenska språket utgiven av Svenska Akademien*. 1898–. Tillgänglig: <saob.se>. Hämtat: maj 2023.

SAOL 8 = *Ordlista över svenska språket utgiven av Svenska Akademien* (8 upplagan, 1923). Stockholm: Svenska Bokförlaget P. A. Norstedt & Söner.

SAOL 9 = *Svenska Akademiens ordlista över svenska språket* (9 upplagan, 1950). Stockholm: Svenska Bokförlaget/Norstedts.

SAOL 10 = *Svenska Akademiens ordlista över svenska språket* (10 upplagan, 1973). Stockholm: Norstedts.

SAOL 11 = *Svenska Akademiens ordlista över svenska språket* (11 upplagan, 1986). Stockholm: Norstedts.

SAOL 13 = *Svenska Akademiens ordlista över svenska språket* (13 upplagan, 2006). Stockholm: Norstedts.

SAOL 14 = *Svenska Akademiens ordlista över svenska språket* (14 upplagan, 2015). Tillgänglig: <svenska.se>. Hämtat: maj 2023.

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

SO 2009 = *Svensk ordbok utgiven av Svenska Akademien* (1 upplagan, 2009). Stockholm: Norstedts i distribution.

SO 2021 = *Svensk ordbok utgiven av Svenska Akademien* (2 upplagan, 2021). Tillgänglig: <svenska.se>. Hämtat: maj 2023.

Språkriktighetsboken 2016. Andra upplagan, fjärde tryckningen. Skrifter utgivna av Svenska språknämnden. Stockholm: Norstedts.

Svensén, Bo 2004. *Handbok i lexikografi. Ordböcker och ordboksarbete i teori och praktik*. 2 uppl. Stockholm: Norstedts.

Svenska skrivregler 2008. 3 uppl. Stockholm: Språkrådet & Liber.

Söderberg, Barbro 1983. *Från rytters och cowboys till tjuvstrykers. S-pluralen i svenskan. En studie i språklig interferens*. Stockholm: Almqvist & Wiksell International.

Wellander, Erik 1970. *Riktig svenska. En handledning i svenska språkets vård*. 3 uppl. Stockholm: Norstedts.

Seddelsamlinger – historisk arkivmateriale eller levende resurse?

Henrik Hovmark

This article investigates the possible, dynamic use and web publication of digitized collections of paper slips – the practical knowledge databases behind a number of often monumental national, historical dictionaries. Using collections at *Ømålsordbogen* (the Dictionary of Danish Insular Dialects) as an example, we demonstrate how an immediate web publication without further introduction will face users, even specialists, with substantial challenges due to the highly densified and specialized form of information on the paper slips, and the sophisticated principles behind the elaboration of the edited entries. However, web publication also has advantages and potentials: useful and otherwise inaccessible information, in the published dictionary entries as well as the underlying database, will become available to specialists as well as common and garden users – provided that carefully selected metadata are added to the paper slips and sufficient general information is put at the disposal of users.

NØGLEORD: leksikografi, seddelsamlinger, digitalisering, digital humaniora, dokumentationsordbøger

1. Indledning

Det er velkendt at leksikografien i de seneste årtier har været igennem omfattende digitaliseringsprocesser som har forandret fagområdet og dets praksis radikalt og fortsat gør det. Et forholdsvis upåagtet hjørne af denne udvikling er digitaliseringen af ældre seddelsamlinger, kildematerialet bag en række, ofte store, nationale ordbøger. Digitalisering i form af tilgængeliggørelse af denne type papirbundne kilder er ikke helt ny i nordisk leksikografisk sammenhæng, jf. pionérarbejdet i Norge med Dokumentationsprojektet og Metaordboken der går helt tilbage til 1990'erne (jf. fx Ore 1995). I Danmark har digitaliseringstypen fået fornyet aktualitet: *Ømålsordbogens* (ØMO) seddelsamling har siden 2018 været genstand for en trinvis digitalisering, og søsterprojektet *Jysk Ordbog* (JO) har udført pilotundersøgelser med henblik på en lignende proces (jf. Hansen et al. 2023).

En digitalisering af en seddelsamling kan i en snæver og intern leksikografisk-redaktionel sammenhæng have det formål at overføre papirbunden information til digital form, med de muligheder og fordele dette har mht. fx søgning og datahåndtering. I en sådan tilgang vil sedlerne så at sige overgå til at være et historisk arkivmateriale i det øjeblik data er blevet overført. Med udgangspunkt i detaljerede analyser af ØMO's seddelsamlinger vil vi i det følgende imidlertid rette fokus mod seddelsamlingernes værdi i deres egen ret og spørge hvordan og i hvor høj grad de kan være eller blive en levende resurse? Hvad kan de bidrage med hvis de tilgængeliggøres i digitaliseret form, typisk på internettet?

Med udgangspunkt i den rolle som ØMO's seddelsamlinger spiller i ordbogens samlinger som helhed og i det leksikografiske arbejde (afsnit 2), vil vi anskueliggøre nogle formidlingsmæssige udfordringer der knytter sig til en umiddelbar tilgængeliggørelse af sedlerne (afsnit 3). Samtidig vil vi give eksempler på de muligheder der også ligger i en digitalisering og offentlig adgang til samlingen (afsnit 4), og afslutningsvis vil vi pege på nogle vigtige erfaringer og fremtidige pejlemærker (afsnit 5).[1]

2. Ømålsordbogens seddelsamlinger og øvrige samlinger

Ømålsordbogen er en af de store nationale, videnskabelige dokumentationsordbøger. Ordbogen beskriver de traditionelle dialekter på Sjælland, Lolland-Falster, Fyn og omliggende øer 1750-1945, med 1850-1920 som kerneperioden. Ordbogen giver fulde, grundvidenskabelige beskrivelser af udtale, bøjning, syntaks, betydning og brug i de enkelte dialekter og dermed også af dansk sprog generelt. Et særkende ved ordbogen er at den giver omfattende oplysninger af etnologisk-encyklopædisk art af den ældre bonde- og fiskerkultur inden for perioden (jf. ordbogens undertitel: "sproglig-*saglig*"). Projektet og indsamlingen startede oprindelig i 1909, men første bind udkom først i 1992. Ordbogen er udkommet regelmæssigt siden, senest bind 12 (lindost-march), og resten af bogstav M er færdigredigeret.

ØMO var oprindelig primært henvendt til forskere og særligt interesserede (fx lokalhistorikere), men ordbogen rummer også mange oplysninger

1 Foredraget ved den 16. konference om leksikografi i Norden berørte også forskellige praktiske vanskeligheder i forbindelse med digitaliseringsprocessen, men de vil af pladshensyn ikke blive berørt i det følgende.

som vil være interessante for en bredere almenhed, og med mulighederne for netpublicering er denne brugergruppe kommet yderligere i fokus (jf. Hovmark 2020). Tilgængeliggørelse i form af digitalisering af de fysiske samlinger føjer sig naturligt til denne type overvejelser. Ømålsordbogens samlinger er imidlertid mangeartede og indholdsmæssigt ofte meget komplekse og specialiserede, og det er langt fra alle der egner sig lige godt til uden videre at blive publiceret på nettet.

Et kort rids af vigtige indsamlingsperioder og -redskaber kan give et indtryk af diversiteten i samlingerne: I 1909-22 anvendtes primært en ordliste ordnet efter lyd og bøjning og en lille blokbog med sedler til spontane iagttagelser. I 1922-35 foregik der et intensivt, stramt organiseret indsamlingsarbejde med brug af billedhæfte, en ordliste ordnet efter kontekst og emne, fx høst, vævning, vejr og vind (Den Store Spørgeliste), fraseologiske spørgelister og fagordsoptegnelser (jf. Gudiksen & Hovmark 2009). Fra 1935 og frem blev der udført kompletterende undersøgelser og indsamlinger af forskellig art. Projektet indsamlede også alt hvad der kunne fremskaffes af materiale der rummede eller mere systematisk beskrev ømålsdialekterne, lige fra ældre bondedagbøger og dialektlitteratur, over lokale dialektbeskrivelser, til videnskabelige monografier, undersøgelser m.v. (ofte arkiveret i Manuskriptsamlingen). Hertil kommer indsamling af lydoptagelser (interview), især i 1970'erne.

Det var – og er – naturligt at relevant information om de enkelte ord samles i én samling før redigering for at gøre processen så effektiv som muligt, på samme måde som en korpusbaseret ordbogsartikel baserer sig på en fremfinding og samlet analyse af ordets forekomster i korpus. Denne opsamling af information på ØMO skete og sker i den alfabetiske samling, og det er denne samling som skal analyseres nærmere i det følgende. Den alfabetiske samling er i sig selv kompleks i sin opbygning, idet den er opstillet i ca. 50 emnesamlinger (fx høst, klæder og syning, atmosfære) foruden en almen restsamling og en accessionssamling – som først samles til én alfabetisk samling i forbindelse med redigeringen. Denne kompleksitet vil vi imidlertid ikke komme nærmere ind på her, men referere til den samlede mængde af ordsedler som ØMO-AlfabetS – uanset om sedlerne er samlet i én samling (ØMO-samlingen, for tiden bogstav A-M), eller om de stadig står spredt i de forskellige emnesamlinger m.v. (for tiden bogstav N-Å).

Sedlerne i ØMO-AlfabetS kan altså beskrives som interne informationsopsamlingsfiler der organiserer oplysninger af meget forskellig type fra mange forskellige kilder. De informationer der står på sedlerne, kan være hentet mere eller mindre direkte fra den talesproglige virkelighed: En redaktionel medarbejder kan have udfyldt sedlerne ved et besøg hos en meddeler eller have renskrevet notater efter besøget; eller lokale meddelere kan selv have udfyldt sedler i besvarelsen af spørgelister e.l. Sedlerne kan også rumme excerperinger fra trykte eller utrykte kilder, fx bondedagbøger, dialektlitteratur – eller monografier om både sproglige og saglige emner. Og ikke mindst vigtigt i det følgende: En meget stor del af sedlerne rummer udskrevet information fra andre af ØMO's delsamlinger, der da får karakter af baggrundssamlinger. Det kan fx være længere, emneorganiserede beskrivelser eller fagordsoptegnelser i Topografisk Samling, eller notater, forarbejder, afhandlinger e.l. i Manuskriptsamlingen. Selvom sedlerne i ØMO-AlfabetS også indgår i et dialektarkiv og som sådan også har en dokumentationsfunktion, er det tydeligt at udformningen af dem er styret af deres interne funktion i redaktionsprocessen. Som det vil fremgå af det følgende, kan dette have både fordele og ulemper.

3. Formidlingsmæssige udfordringer i forbindelse med den alfabetiske seddelsamling

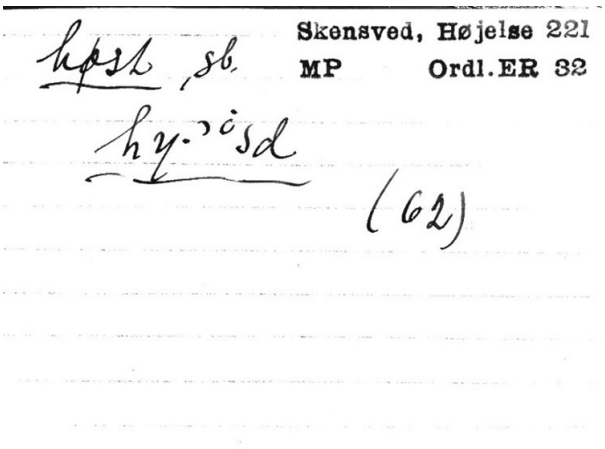
En digitalisering af ØMO-AlfabetS og ordbogens øvrige (seddel)samlinger vil give mulighed for at ordbogsbrugerne kan få adgang til ordbogsdata – og dermed redaktionens valg. Med andre ord vil man få mulighed for at sammenholde den redigerede ordbogstekst med kildematerialet eller om man vil: kigge redaktørernes valg efter i sømmene. Dette er principielt rigtigt, men i praksis er den blotte adgang til materialet ikke nødvendigvis ensbetydende med (større) indsigt i og dermed mulighed for at vurdere de analyser og valg som redaktionen har gjort. En række forhold gør det vanskeligt overhovedet at afkode hvad der står på sedlerne, ligesom tolkningen af indholdet kræver stor forhåndsviden og vil være vanskelig, selv for andre specialister.

Man kan skelne mellem to trin eller processer i vejen fra den sproglige virkelighed og frem til den leksikografiske repræsentation af denne virkelighed i en ordbog (jf. Hovmark 2012):

1) Indsamling og udvælgelse af repræsentative data fra virkeligheden > ordbogsdata (i form af fx digitale tekstkorpora eller seddelsamlinger).
2) Analyse, organisering, redigering og præsentation af ordbogsdata > ordbogsartikel (eller anden form for præsentation).

Ordbogssedlerne i ØMO-AlfabetS rummer information om begge trin, som hver især involverer forskellige vurderinger og valg (ligesom det vil være tilfældet når ordbogsdata består af digitale tekstkorpora). Hvad angår trin 1, vil sedlerne rumme objektiv information, fx om hvor en informant kan lokaliseres ('sted'), men også vurderinger og valg, fx en vurdering af hvordan en informant har udtalt et ord, og et valg af en bestemt notation af den pågældende udtale. Tilsvarende vil trin 2 ikke kun bestå af simpel organisering og videreformidling af sikre og utvetydige fakta (om køn, bøjning, betydning osv.) – processen vil også omfatte vurderinger og tolkninger, især når materialet er så heterogent i både tid, form og indhold som det er tilfældet med ØMO's samlinger. Det kan fx være vurderinger af hvordan en udtale nedfældet i en grov eller ældre lydskrift skal tolkes, eller hvordan en afvigende udtale nedfældet i en fin lydskrift af en i øvrigt pålidelig optegner skal behandles. At vurdere en redaktørs eller redaktions brug af et kildemateriale (ordbogsdata) er med andre ord ikke en enkel opgave. Lad os i det følgende se nærmere på de udfordringer som ØMO-AlfabetS specifikt stiller brugeren over for.

Den første og måske største forhindring består i at de fleste oplysninger på netop sedlerne i ØMO-AlfabetS er stærkt kodede. Dette hænger direkte sammen med sedlernes status og funktion som dataopsamlingsfil i den interne informationsorganisering (jf. afsnit 2). Mange informationer er forkortede, af både tids- og pladsmæssige grunde. Det er hurtigere at skrive forkortelsen PA (ØMO's mangeårige leder Poul Andersen) end hele navnet, og forkortelsen kan bekvemt passes ind i øverste højre hjørne sammen med andre basisoplysninger. Dette ses fx på den seddel der er afbildet på Figur 1.



FIGUR 1. Ordbogsseddel med information om udtalen af ordet *høst* i sognet Højelse på Sjælland.

Sat på spidsen rummer sedlen på Figur 1 kun én indholdsoplysning om den sproglige virkelighed, nemlig en udtaleoplysning (af ordet *høst*, i Danias lydskrift, i midten). Alle øvrige oplysninger er metadata, og de fleste er anført i en form som ikke er umiddelbart forståelig for andre end redaktørerne. "Skensved, Højelse 221" er geodataoplysninger, obligatoriske for beskrivelserne i en ordbog over geografisk relateret sproglig variation. Selvom stednavnene ikke vil være almindeligt kendte (landsbyen (Lille) Skensved i Højelse sogn på Sjælland ved Køge bugt), er oplysningerne dog ikke kodede, bortset fra tallet 221 der refererer til standardsognenummeret for Højelse sogn. De resterende oplysninger er til gengæld stærkt kodede: "MP" = informantinitialer (uden vurdering); "Ordl." = ordliste transskriberet af "ER" = redaktøren Ellen Raae i "32" = 1932; og det håndskrevne "(62)" = nummeret på det spørgsmål i ordlisten som sedlens data svarer på: "ø foran st", dvs. vokalen ø før konsonantkombinationen -st.

Disse oplysninger er interne basisoplysninger, metadata, der hver især kan vise sig vigtige i trin 2 i redaktionsprocessen, nemlig når alle sedler med ordet *høst* samles af de nuværende redaktører og analyseres i forbindelse med udarbejdelsen af ordbogsartiklen. Her kan det vise sig meget relevant fx at vide hvem kilden er, hvem redaktøren er, og under hvilke omstændigheder udtaleoplysningen er blevet indhentet.

Sedlen rummer dog ikke kun interne forkortelser og information. Forkortelsen "sb." vil fx være velkendt, i hvert fald for sprogforskere og trænede ordbogsbrugere. Lydskriften er nedfældet i lydskriften dania, som også vil være kendt blandt nogle sprogforskere. Alligevel vil afkodningen af udtaleoplysningerne i ØMO-AlfabetS være vanskelig. En række forskellige lydskrifter af mere eller mindre systematisk art vil optræde på sedlerne, og selv udtaler i dania skal læses kritisk idet der fx kan være forskellige optegnervaner på spil. Det er derfor nødvendigt at foretage en tolkning og generalisering af oplysningerne med henblik på en blot nogenlunde overskuelig og sammenlignelig fremstilling i ordbogsartiklen. Figur 2 viser udtaleafsnittet i artiklen *høst*.

høst s m M, L-Fa; m og f F, T, Lgl, dog > c S, c Æ
[*hɥ'ɜsd*] nS(vist alm), nvS(Ods hrd alm, Bregn, SJern, LFugl, Røs, Svall, KHels, tilsv Sejr Thorsen.SS:61), svS (Ørs), øS(alm, jf FØS:345), øF(alm, jf FØF:229), u stød svS(Agrs, Omø), ssS, sF, T, Lgl, Æ(alm), [*høɜsd*] M, L-Fa (alm), [*hoj'sd, hȯj'sd*] vF(alm, tilsv PJac.MF:81$_2$); [*høsd, hösd*] Am(Drag), nS(vist alm), vS(alm undt Ods hrd), sS (Bår, Ever); komm: om grænsen ml *ɥɜ* og *ȯj, oj* på Fyn, se ØMO.Till: kort 22. – Endv: bf sg: [*hɥɜsdni*] T(Land).

FIGUR 2. Udtaleafsnittet i artiklen *høst*.

Udtaleafsnittet til *høst* kan forekomme kompliceret, men i virkeligheden gengiver dette lydhoved kun forventede, regelrette former og udtaler af netop denne ordtype i de forskellige dialektale hovedområder i ømålsområdet – bortset fra den sidste form i afsnittet indledt med "Endv". Afsnittet hjælper altså brugeren ved at analysere og kondensere det brogede kildemateriale på videnskabeligt grundlag: Upålidelige belæg er siet fra, og ikke-betydningsbærende variation er blevet generaliseret. En tilgængeliggørelse af seddelmaterialet og dets oplysninger om lydlige former vil på den ene side give mulighed for at vurdere redaktionens analyser og generaliseringer. På den anden side er principperne for udarbejdelsen af udtaleoplysningerne i trin 2 meget komplicerede, selv for specialister, idet de enkelte lydformer på samme tid søger at gengive både fonetisk form og fonematisk struktur. En umiddelbar tilgængeliggørelse af sedlernes udtaleoplysninger vil derfor kunne skabe større forvirring hvis de mange mel-

lemregninger der indgår i afkodning og tolkning af oplysningerne ikke samtidig stilles til rådighed. Andre oplysninger på sedlerne kan imidlertid være lettere at afkode og har potentiale til at fungere som et nyttigt supplement til ordbogsartikel og ordbogsbase.

4. Den alfabetiske seddelsamling som forsknings- og formidlingsresurse

Den redigerede ordbogsartikel kan beskrives som en koncentreret repræsentation af ordbogsdata. Koncentratet vil være kvalitativt og, som vist, afspejle forskellige faglige, videnskabelige valg og vurderinger. Men koncentratet vil også være kvantitativt i en mere simpel forstand: Ordbogsartiklen kan og vil normalt kun bringe en (lille) delmængde af ordbogsdata. Nyere, korpusbaserede ordbøger kan her supplere den redigerede repræsentation med adgang til de konkordanslinjer som har dannet baggrund for en given artikel og fx give langt flere eksempler på ords betydninger og kontekstuelle brug end dem der er udvalgt som repræsentative i fx citater, kollokationsmønstre e.l.

En tilgængeliggørelse af digitaliserede ordbogssedler har principielt samme potentiale. Et ideelt eksempel i ØMO er artiklen *kromand* (jf. Hovmark 2012). Denne artikel bringer to citater, men et kig i seddelsamlingen afslører yderligere syv sedler med fraseologisk materiale, to sedler med substantivsyntagmer samt to ordsprog som redaktøren har udeladt. En tilgængeliggørelse kan altså give indblik i redaktørernes valg af citater, herunder om eller i hvilken grad citaterne gengiver typiske kontekster eller konstruktioner. En tilgængeliggørelse kan imidlertid også bidrage til at det fraseologiske materiale i ØMO's samlinger generelt kan få større opmærksomhed og udnyttelse. Netop dette materiale står stærkt i ØMO's samlinger, bl.a. takket være målrettede indsamlinger (jf. afsnit 2).

Citatmaterialet i ØMO-AlfabetS er også en oplysningstype som kunne blive til glæde også for ikke-specialister og mere alment interesserede brugere, ikke mindst hvis det kan kobles til de enkelte (under)betydninger i en netudgave. En umiddelbar tilgængeliggørelse er dog heller ikke uden vanskeligheder her. Mange sprogeksempler er nedfældet i lydskrift eller i mere eller mindre (u)tydelige håndskrifter, og der vil stadig være udfordringer med at forstå de kodede oplysninger på sedlen. Hertil kommer at det fraseologiske materiale, i lighed med betydningsbeskrivelser i det hele taget,

kan være meget begrænset, og bl.a. derfor kan også betydningsbeskrivelser involvere temmelig komplicerede redaktionelle analyser og vurderinger. Et mindre, men meget værdifuldt korpus af lydoptagelser udskrevet i standarddansk giver dog adgang til citatmateriale som er lettere at afkode (jf. Gudiksen & Hovmark 2008).

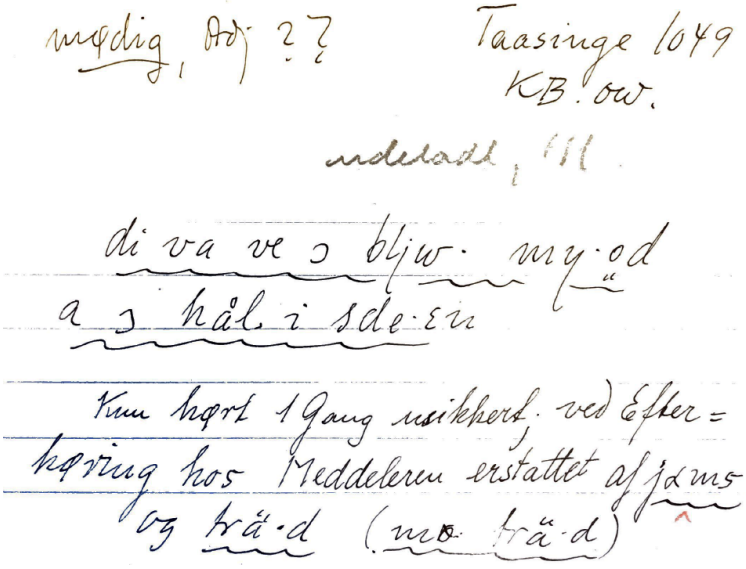
Et sted hvor seddelmaterialet rummer mere information end ordbogsartiklen, er ved geodataoplysninger. Det er almindeligt at generalisere ords eller betydningers udbredelse, til fx S(spor opt) 'dialektområdet Sjælland sporadisk optegnet og muligvis almindeligt', Ø(alm) 'dialektområdet ømålsområdet almindeligt udbredt' m.fl. I disse tilfælde vises de specifikke lokaliseringer, som oftest sogne men også fx herreder, egne e.l., ikke i den redigerede artikel. Ordet *kromand* er fx beskrevet som "Ø(alm)" – og neden under denne generalisering gemmer sig i alt 39 belæg fra generelt sikre kilder jævnt fordelt i hele ømålsområdet.

De specifikke lokaliseringer går imidlertid ikke kun tabt i ordbogsartiklen – ordbogsdatabasen er i øjeblikket indrettet sådan at de heller ikke registreres her hvis der generaliseres. Det betyder at det ikke er muligt at søge målrettet og/eller udtømmende på specifikke lokaliteter på sogneniveau e.l. En tilgængeliggørelse af ØMO-AlfabetS vil afhjælpe dette problem da alle lokaliseringer på sedlerne vil blive synlige. En søgning på geodata vil dog først blive fuldstændig og effektiv hvis sedlerne forsynes med metadata om 'geodata/sted' i en base, i tilgift til den primære metadataoplysning 'opslagsord'. En nylig forskningsundersøgelse illustrerer både problemet og potentialet. Med henblik på at afdække eventuelle sproglige særtræk i dialekterne ud mod Øresund og evt. tilgrænsende områder i Nordsjælland var det relevant at kunne finde belæg optegnet i specifikke sogne (Hovmark 2021). Her gav en søgning i ordbogsdatabasen imidlertid kun resultat i de tilfælde hvor de relevante sogne ikke indgik i en generaliseret udbredelse, hvilket gav et alt for tilfældigt billede til at kunne sige noget samlet om de sproglige forhold.

Tilgængeliggørelse af ØMO-AlfabetS vil altså give mulighed for nye måder at udnytte de oplysninger og den viden på som ligger i ØMO-samlingerne som helhed, som et supplement til de oplysninger der allerede er søgbare i ordbogsdatabasen og repræsenteret i ordbogsartiklerne. En forudsætning er dog at der tilføjes flere centrale metadata. I den ovennævnte forskningsundersøgelse vidste man fx fra andre kilder at der var blevet udført optegnelser af bestemte, navngivne medarbejdere, og hvis 'opteg-

ner' e.l. også var tilføjet som metadata, ville man kunne udføre endnu mere præcise søgninger.

Det er imidlertid ikke alt i ØMO-AlfabetS der er svært forståeligt eller underforstået. Netop fordi sedlerne er levende redaktionsopsamlingsfiler, rummer de i visse tilfælde også ekspliciteringer af de valg og vurderinger som redaktørerne har gjort, fx bemærkninger om hvordan en oplysning eller en kilde måske (ikke) bør tolkes eller kan tolkes som.



FIGUR 3. Seddel med kritiske redaktørkommentarer til et muligt belæg på ordet *mødig* 'træt'.

Sedlen på Figur 3 giver eksempler på denne type information. Den oprindelige optegner, OW (Ole Widding), har anført alvorlige kildekritiske bemærkninger til et belæg på adjektivet *mødig* i betydningen 'træt': "Kun hørt 1 Gang usikkert, ved Efterhøring hos Meddeleren erstattet af *hjams* [uoplagt, mat, sløj] og *træt*". Bemærkningen er formodentlig anført i tilknytning til trin 1, men med tanke på det senere trin 2: Der er sat hele to spørgsmålstegn efter opslagsordet, som udtryk for at sedlens oplysning om en sproglig virkelighed i dette tilfælde anses for tvivlsom. Den nuværende redaktør har valgt at følge OW's indstilling og tilføjet: "udeladt, HH". Dels er OW generelt en både sikker og kritisk optegner, dels viste det sig at belægget på *mødig* stod alene på Tåsinge.

Som det fremgår, indgår disse bemærkninger i den interne, redaktionelle informationopsamling og -udveksling, men oplysningerne, der meget sjældent vil være tilgængelige eller opsamlet overhovedet, heller ikke i moderne korpusbaserede ordbøger, kan være interessante for andre, eksterne brugere.

5. Sammenfatning og fremtidige pejlemærker

Undersøgelsen her har vist at sedlerne i ØMO-AlfabetS rummer et stort antal interessante oplysninger. Undersøgelsen har imidlertid også vist at en umiddelbar tilgængeliggørelse, uden forskellige former for vejledning, vil være utilfredsstillende for stort set alle brugere: Oplysningerne på sedlerne er stærkt kodede, og principperne for redaktionens valg og vurderinger vil ofte basere sig på komplicerede forudsætninger som ikke fremgår af sedlerne. Hertil kommer at mange af sedlernes oplysninger allerede er tilgængelige i langt mere forståelig form i den redigerede ordbog og i ordbogsbasen – som netop er resultatet af en nøje faglig og forskningsbaseret vurdering og generalisering.

Undersøgelsen har imidlertid også vist at sedlerne rummer oplysninger der ikke fremgår af den redigerede artikel og undertiden heller ikke er opsamlet i ordbogsbasen. Det kan derfor alligevel være relevant at overveje en tilgængeliggørelse. Spørgsmålet er blot hvordan? Forskellige former for formidlende tiltag vil være relevante, men generelt vil det også være nødvendigt at tilføje metadata til sedlerne. Samtidig viser eksemplerne at det er en god idé at overveje hvilke metadata det vil være frugtbart at investere i: Hvor rummer sedlerne særlig værdifuldt indhold? Det kan som vist fx være geodataoplysninger, hvilket vil styrke forskningsmulighederne i ØMO som helhed, eller det kan være betydningsnumre så forskellige brugere ud fra både et formidlings- og dokumentationshensyn kan få bedre adgang til kildematerialet.

Der findes utvivlsomt andre seddelsamlinger som er mindre heterogene end ØMO-AlfabetS i både form og indhold, og som derfor egner sig bedre til en umiddelbar publicering. Alligevel peger undersøgelsen her på at det måske vil være hensigtsmæssigt i alle tilfælde at overveje hvordan en given seddelsamling mere specifikt kan udfylde en meningsfuld funktion i samspil med øvrige resurser på en digitaliseret ordbogs platform. Men det forudsætter at man anerkender og har blik for seddelsamlingen som en

informationsresurse i egen ret og ikke blot som et historisk vedhæng til et moderne publiceret produkt.

Litteratur

Gudiksen, Asgerd & Henrik Hovmark 2009. Måske husker De noget alle andre har glemt. I: Gudiksen, Asgerd, Henrik Hovmark, Pia Quist, Jann Scheuer & Iben Stampe Sletten (red.), *Dialektforskning i 100 år*. København: Københavns Universitet, 13-64.

Gudiksen, Asgerd & Henrik Hovmark 2008. Båndoptagelser som kilde til Ømålsordbogen. I: Svavarsdóttir, Ásta, Guðrún Kvaran, Gunnlaugur Ingólfsson & Jón Hilmar Jónsson (red.), *Nordiske Studier i Leksikografi* 9. Reykjavík: Nordisk Forening for Leksikografi, 173-182.

Hansen, Inger Schoonderbeek, Mette-Marie Møller Svendsen & Kristoffer Friis Bøgh 2023. Digitalisering af Jysk Ordbogs seddelsamling. I: Holmer, Louise et al. (red.), *Nordiska studier i lexikografi 16*. Lund & Göteborg: Nordiska föreningen för lexikografi, 99-112.

Hovmark, Henrik 2012. Data og repræsentativitet i ordbogsarbejdet. I: Eaker, Birgit, Lennart Larsson & Anki Mattisson (red.), *Nordiska studier i lexikografi* 11. Lund: Nordisk förening för lexikografi, 296-308.

Hovmark, Henrik 2020. Ømålsordbogens brugere – nu og i fremtiden. I: Sandström, Caroline, Ulla-Maija Forsberg, Charlotta af Hällström-Reijonen, Maria Lehtonen & Klaas Ruppel (red.), *Nordiska studier i lexikografi* 15. Helsingfors: Nordisk förening för lexikografi, 145-153.

Hovmark, Henrik 2021. Dialektoptegnelser fra øresundsfiskerlejerne i 1932- og spørgsmålet om øresundsmålene. *Danske Talesprog* 21, 29-55.

JO = *Jysk Ordbog*. <www.jyskordbog.dk>. Hentet november 2022.

Ore, Christian-Emil 1995. Korpus og seddelarkiv, fredelig sameksistens mellom det beste og det gode? I: Svavarsdóttir, Ásta, Guðrun Kvaran & Jón Hilmar Jónsson (red.), *Nordiske Studier i Leksikografi* 3. Reykjavík: Nordisk Forening for Leksikografi, 331-338.

ØMO = *Ømålsordbogen. En sproglig-saglig ordbog over dialekterne på Sjælland, Lolland-Falster, Fyn og omliggende øer* 1-. 1992-. København: Universitets-Jubilæets danske Samfund.

Facebook som kjelde for norske dialektord. Verdifullt nytt tilfang for *Norsk Ordbok*

Tor Erik Jenstad

This paper will discuss Facebook groups as a source for new information about the vocabulary in Norwegian dialects, especially for collection of material relevant for inclusion in the dictionary *Norsk Ordbok*. A short introduction is given to this dictionary and its history, and also a short overview of the collection of Norwegian dialect words until today. The paper will give examples of material being brought forth through these Facebook groups, and how this can be used in our editorial work in the ongoing revision project NO-AH. The paper will critically examine Facebook as a source and discuss both its possible drawbacks and benefits.

NØKKELORD: leksikografi, Norsk Ordbok, dialektord

1. Innleiing

Artikkelen tek for seg Facebook som kjelde til ny informasjon om ordforrådet i norske dialekter, særleg med tanke på bruk i det redaksjonelle arbeidet i *Norsk Ordbok* A-H. Det blir gitt eksempel på relevante opplysningar som er komme inn gjennom Facebook-gruppene. Til slutt blir Facebook som kjelde drøfta kritisk, med fordelar og ulemper.

2. Norsk Ordbok A-H

Forfattaren arbeider på prosjektet NO A-H (forkorta for Norsk Ordbok frå A til H). Norsk Ordbok på papir vart fullført med tolv band i 2016. Første bandet kom i 1966. Norsk Ordbok skal dekke norske dialektar og det nynorske skriftspråket, eller «Ordbok over det norske folkemålet og det nynorske skriftspråket», som det står i undertittelen både i papirutgåva og på nettsidene. Norsk Ordbok er nok velkjent for mange av lesarane, bl a gjennom fleire presentasjonar på NFL-konferansane (sjå t.d. Ore & Wetås 2014). Om Norsk Ordbok sjå også Urdland Karlsen et al. (2016).

Dei siste sju banda, frå I til Å, er også publisert digitalt. I prosjektet NO A-H skal dei fem første banda digitaliserast, reviderast og oppgraderast til same standard som dei sju siste banda. Prosjektet er eit samarbeid mellom Universitetet i Bergen og Høgskulen i Volda. Det vart starta i 2019 og er finansiert gjennom årlege løyvingar frå det norske Kulturdepartementet.

I denne revisjonen av første del av ordboka må det også leggast til dialekttilfang. Det kan gjelde nye ord (og sjølvsagt også «gamle» ord som ikkje tidlegare er registrert), nye uttaleopplysningar, bøyingsformer og tydingar, altså informasjon i alle delar av ein ordboksartikkel. Samtidig må ein supplere informasjon om geografisk utbreiing på alle desse forskjellige nivåa. Det ligg ei utfordring i å utfylle og oppdatere dialektmaterialet, når det ikkje lenger blir drive aktiv ny innsamling i offentleg regi. Norsk Målførearkiv er lagt ned for lenge sidan[1], og ved Språksamlingane i Bergen har ein ikkje ressursar til anna enn å ta imot og registrere det ein får inn.

3. Registrering av ordtilfanget i norske dialektar

Det som kan kallast ei systematisk registrering av ordforrådet i norske dialektar begynte på 1800-talet, med ordbøkene til Ivar Aasen (1850, 1873) og Hans Ross (1895). Før dette hadde det vore meir sporadiske tilfelle, heilt tilbake til 1600-talet. Somme av desse var riktignok ganske omfattande, som Hallager 1802 (Norsk Ordsamling) og Christie 1937 (Norsk Dialect-Lexicon; med manuskript frå 1830–40, publisert først 1937). Aasen samla dialektord som tilfang til landsmålet. Ross (1895) var tenkt som eit tillegg til Aasen, slik det også står i forordet, men den endte med å bli større. Dette har sjølvsagt å gjera med metoden til Ross: Han gjekk mykje meir i detalj, noterte variantar og særeigenheiter, utan å ha som direkte formål at dette skulle inn i noko standard skriftspråk.

Prosjektet Norsk Ordbok kom i gang rundt 1930. Det vart verva medarbeidarar frå rundt i heile landet (Vikør & Wetås 2016). Desse skreiv

1 Ved Norsk Målførearkiv (1936–1990) vart det mellom anna sendt ut spørjelister om nemningsbruk til eit nett av medarbeidarar. Etter 1990 gjekk målførearkivet inn i Seksjon for leksikografi og målføregransking ved Institutt for nordistikk og litteraturvitskap ved Universitetet i Oslo. I 2016 vart alt tilfanget flytta til Språksamlingane ved Universitetet i Bergen. Enno sist på 1990-talet var medarbeidarnettet såpass intakt at artikkelforfattaren kunne bruke det.

ned dialektord på setlar og sendte dei inn til redaksjonen. På det meste var det meir enn 1000 slike medarbeidarar (sjølv har eg medarbeidarnummer 1022, og vart med frå ca. 1987). Dette innsamla materialet varierer veldig, både i omfang, ambisjonsnivå og kvalitet. Dei arbeidde på forskjellige måtar: Somme gjekk systematisk gjennom Aasens ordbok og noterte korleis orda og eksempla ville bli brukt på eigen dialekt. Andre kunne skrive ned sitat frå bøker (som dei gjerne fekk tilsendt frå redaksjonen, med dei aktuelle orda understreka) og føre på same setel korleis dette ville bli uttrykt på eigen dialekt. Alt i alt er dette eit veldig stort materiale om ordforrådet i norske dialektar, og det er tilgjengeleg digitalt under grunnlagsmaterialet til Norsk Ordbok (Setelarkivet og Metaordboka, med om lag 3,3 millionar belegg; Norsk ordbok a).

Så har det oppgjennom åra vore publisert mange ordsamlingar for enkeltdialektar (pluss nokre, men ikkje så mange, med eit regionalt siktemål), og dette har auka sterkt dei siste åra. Dette er sjølvsagt eit svært verdifullt materiale for Norsk Ordbok, og det må seiast å vera ein framifrå prestasjon og stor arbeidsinnsats, i all hovudsak gjort av amatørar.

Elles kan ein nemne systematiske kartleggingar av ordforrådet på visse avgrensa felt, i ordgeografiske arbeid, som Bandle 1967, Aune 1976, Jenstad 2001 og fleire mindre arbeid av Arnold Dalen (t.d. Dalen 1967, 1985).

Dermed skulle ein tru at det norske dialektordforrådet var temmeleg grundig dokumentert, frå dei første leksikografane som arbeidde systematisk med dette (Aasen 1850, 1873; Ross 1895) og fram til dagens eksplosjon av lokale ordsamlingar. Men når ein går gjennom dei ulike kjeldene som er tilgjengelege, ser ein fort at det faktisk er betydelege lakuner, både i det tradisjonelle ordforrådet og i neologismar som ventar på å bli oppdaga. Standardisering av ordforrådet er jo ein hovudtendens i norsk talemålsutvikling i dag. Men trass i dette foregår det stadig ein viss lokal og regional eigenproduksjon. I andre land, som Sverige og Danmark, får ein det inntrykket at å samle dialektord meir eller mindre hører fortida til.

Dette har å gjera med dialektsituasjonen i Norge. Også desse er i sterk endring, kanskje særleg i ordforrådet, som nemnt, men dei er likevel i høg grad levande og faktisk i aukande bruk i det offentlege rommet og i digitale media. Det er sterk interesse for dialektar, noko som avspeglar seg mellom anna i dei mange lokale og regionale ordbøkene som kjem i ein

stadig straum[2], og ikkje minst nettsider for dialekt, eit forum som også er i sterk vekst. Fleire av dei lokale ordsamlingane som kjem, blir no (berre) publisert på nettet (sjå t.d. Vallemål 2019 og Numedalsmål 2023).

4. Korleis registrere nytt materiale

No for tida er sjølvsagt all vitskapleg leksikografi korpusbasert. Neologismar blir i all hovudsak også registrert på denne måten, gjennom skrivne eller talemålsbaserte korpora. Andre metodar som har vore brukt, er opptak, intervju og spørjelister. Men nyleg har nettet i aukande grad vorte brukt for å samle lingvistiske data, i ei form for crowd-sourcing. For Norge sin del kunne ein t.d. nemne Opsahl 2015, der barn og ungdom skulle registrere ord som "dei vaksne ikkje kjenner». Vi har også kampanjen "Ta tempen på språket» frå 2014, der elevar frå heile landet skulle samle data om deira eige språk, også ordforråd.

Kanskje den mest pålitelege metoden for å samle levande ordforråd ville vera deltakande observasjon, altså eit slags antropologisk metode (som eg har skrive om i Jenstad 2011), men dette er tidkrevjande og oftast ikkje praktisk mogleg.

Det ser ut til å vera gjort lite forskning på å hente ordforråd frå Facebook (i det minste når det gjeld norske forhold). Eg har sjølv skrive ein liten artikkel om det i ein lokal publikasjon (Jenstad 2012), da Facebook-grupper med dialektfokus for alvor begynte å blomstre opp.

5. Facebook-gruppene

Det er ein blømande aktivitet i dei etter kvart nokså mange Facebook-gruppene som fokuserer på dialektar i Norge. Desse er i hovudsak av to slag: Den eine typen relaterer seg eksplisitt til dialektar, men titlar som «ord og uttrykk frå», «dialektord frå». Den andre gjeld historia til eit område, der dialekt ofte kjem opp som eit naturleg diskusjonstema i trådane, som «Du veit du er frå … når …» («når du seier … xx»; «når du veit kva …. betyr»). Mange av gruppene er nokså lokale, for ei bygd eller ein kommune, medan andre dekkjer eit større område, som Trøndelag, Sunnmøre og Nordfjord. Somme av dei kan ha opp til fleire tusen medlemmer.

2 Eit par døme på fyldige, lokale ordsamlingar er Donali & Indseth (2015) og By (2022).

Dialekt er ein viktig del av den sterke lokale identiteten i Norge, og begge typane grupper viser sterk entusiasme og stort engasjement. Typisk startar ein slik tråd med at eit medlem postar eit ord eller uttrykk (med eller utan kontekst, det første fungerer sjølvsagt best). Så fører dette ofte til lange diskusjonar og mange kommentarar, ofte er desse også skrivne på dialekt og gir såleis ekstra materiale. Det kan vera usemje om kva som er korrekt dialektbruk (her kan det vera sterke meiningar og kategoriske uttalelsar og påstandar «dette har eg aldri hørt», «nei, det har vi aldri sagt her»), det kan komma for ein dag intern variasjon i det aktuelle språksamfunnet, og det er spørsmål om opphav/etymologi og geografisk distribusjon/utbreiing. Redaksjonen i NO A-H kan ofte få gode sitat (eksempel på bruken) og gode historier som er knytt til orda som er under diskusjon.

Somme medlemmer deltek i og bidreg til fleire grupper, ofte i nærliggande område. Stundom blir det sett opp kvissar og undersøkingar for å kaste lys over bestemte emne. Dette skjer gjerne som eit slags meiningsmålingar der det blir sett opp alternativ, og så skal ein krysse av kva ein brukar sjølv.

Somme bidrar med lange historier på dialekt, eller dei deler eldre manuskript, tekstar dei ønsker å få kommentert. I somme grupper blir det innkomne materialet samla til større dokument, helst ved administratorane.

Aktiviteten varierer sjølvsagt både frå gruppe til gruppe og internt i gruppene. Det kan vera rolege periodar der det ikkje skjer så mykje, og så bryt det plutseleg laus igjen. Ein ting som rett nok ikkje er undersøkt, men som ein nok kan vera rimeleg sikker på, er at fleirtalet av medlemmene nok er ganske godt opp i åra. Det kjennest viktig å få dokumentert ein dialekt som er i stadig endring før den eventuelt forsvinn heilt. Språket folk lærte i barndommen står i eit spesielt, gunstig lys, som det beste og det korrekte, og dette gjennomsyrer aktiviteten i gruppene. Så ein må nok innrømme at nostalgi er ei viktig drivkraft.

Det fanst ei tid ei gruppe for «Utrydningstruede ord og uttrykk». Der diskuterte ein både ord frå standardspråket og dialektord som er i ferd med å gå av bruk. Naturleg nok var det her ein stadig pågåande diskusjon om kva som faktisk var verdt å nemne i ei slik gruppe. Gruppa er no lagt ned. Nedlagte grupper blir gjerne arkivert, slik at det stadig går an å leite dei opp og søke i innlegga.

Eg er sjølv medlem eller følgjar i ganske mange av desse gruppene. Ofte blir eg invitert inn, for å svara på spørsmål og komma med faglege kom-

mentarar. Eg brukar Facebook aktivt for å hente informasjon om dialektalt ordforråd. Eg har drive med dette sidan om lag 2011–2012 og fram til dags dato, altså godt og vel ein tiårsperiode. Kor mange slike grupper som faktisk eksisterer, er nesten uråd å seie. Dette er eit dynamisk fenomen, og det kan variere nesten frå dag til dag. Men eg er kjent med om lag 60 slike grupper i alt. Det er utan tvil mange fleire.

I det følgjande vil eg gi nokre eksempel på leksikalsk materiale som er komme inn gjennom desse Facebook-gruppene, og korleis det kan brukast i det redaksjonelle arbeidet i prosjektet NO A-H.

6. Eksempel og resultat

Eksempla er tekne frå bokstaven A, der redigeringsarbeidet i prosjektet NO A-H har begynt. Eg baserer meg på ei oppteljing som vart gjort ved slutten av vinteren 2021. Det har komme til meir materiale etter det, men det skulle likevel gi eit inntrykk.

Det er ny informasjon om til saman 170 ord (lemma). 43 av desse var ikkje tidlegare registrerte, og er derfor kandidatar til nye lemma i NO A-H. Av desse er 31 substantiv (30 samansetningar, 1 avleiing), 10 adjektiv (9 samansetningar, 1 avleiing) og 2 verb (1 samansetning, 1 avleiing). At samansetningane dominerer, er som ein vil vente, sidan dette jo er den viktigaste ordlagingsmåten i germanske språk (Fjeld & Vikør 2008:52). Det er altså ikkje komme informasjon om nye grunnord. Men det er komme relevant informasjon om 40 grunnord (mest substantiv, deretter verb, men også 2 adjektiv, 2 adverb, 2 preposisjonar, 2 interjeksjonar og 1 pronomen). Nye uttaleformer er registrert for 11 ord, ny informasjon om geografisk utbreiing for 49 ord, og nye uttrykksvariantar for 16 ord.

Som døme på ei ny samansetning kan vi ta *aprilhagl*, i ordtaket «aprilhaggel e jøssel te bonins åker» (Kolvereid i Namdalen, posta 24.5.2016). Dette kan det vera aktuelt å føre opp som lemma fordi det går inn i ei fast formulering, men redigeringa er ikkje komme så langt enno.

Døme på eit nytt verb kan vera *annast* – 'når e ha verre bort i ri, anas e hematt, stunde, lengte' (Rennebu). Enten er dette ein variant av det refleksive verbet *annast* som er ført opp i Norsk Ordbok under *III anna* (I: 105), eller ei avleiing *anast* med tilknytning til *I ana* '(om dyr) taka vêret, vêra; gå, stå spanande, uroleg, stundande' (Norsk Ordbok I:73) . I alle fall er

denne bruken ikkje tidlegare belagt. Ordet er også komme med i ei ny ordsamling frå Rennebu (Fjellstad 2020:14).

Døme på ny uttaleform av grunnord: *attjø* (Surnadal, Rindal, Åsskard), som er redigert inn i artikkelen *II adjø* (Norsk ordbok b).

Facebook kan brukast aktivt for å innhente informasjon om bestemte ord. Eg posta eit spørsmål om uttalen av ordet *aksje* i to dialektgrupper. Den eine er kalla «Dialektord» og skal i prinsippet dekke heile landet, men det er nok eit klart tyngdepunkt sønnafjells, og den andre er «Trønderske ord og uttrykk som burde brukes oftere». Som eit resultat kunne uttaleforma *aksji* leggast til for 11 nye kommunar, og *aksi* for fire. Dette kan kanskje synast som bagatellar, men det gir oss eit betre totalbilde av ordet, og i dette tilfelle av uttalen.

Opplysningane blir lagt inn på elektroniske setlar i arkivet til NO, og blir såleis tilgjengelege også utanom redaksjonen.

Men no skal vi ein tur ut i kornåkeren, og samansetningar med ordet *agn*, som kan brukast om dekkblad og snerpe. Ofte er det brukt i fleirtal, jf. svensk *agnar*, dansk *avne(r)*.

Det har to tydingar: Botanisk gjeld det 'dekkblad rundt småaks og blomster på grasplanter'. Den andre og meir «folkelege» tydinga er 'busta på kornakset eller enkelt korn' (ofte brukt samla om begge).

Eg var ute etter samansetningar med *agn-* og posta på Facebook. Her bør det seiast at Norsk Ordbok har eit spesielt ansvar for å dekke ordforråd knytt til det seine bondesamfunnet og overgangen til eit tidleg industrielt samfunn. I treskinga var det ein person som hadde arbeidet med å fjerne, ta unna agnene. Oftast var det ein gut, og denne jobben var på botnen av hierarkiet, han var lågast på rangstigen. Men det kunne også vera kårkallen. Det var fælt, i føyka av støv og snerpe. Nye ord (som vi ikkje hadde frå før) for dette er *agnefis*, *agnegut* og *agnekusk*, og utbreiingsområdet for *agnekuse* er utvida. For reiskap i samband med handteringa av agnene har vi fått inn *agneblåse* – ei vifte – og *agneskoke* (*agnskokko*–Overhalla), som også er eit sorteringsreiskap. For stader for å lagre agnene har vi fått inn *agnestål* i tillegg til dei vi hadde frå før (som *agnebinge*, *agnebu*, *agnehus*).

Alt i alt ser vi at vi gjennom dette har fått ei mykje betre dekning av ordforrådet på dette feltet. Samtidig viser det at langt frå alt er innsamla, sjølv ikkje ordforråd knytt til tradisjonelt arbeidsliv.

Til slutt har vi eit eksempel på utvida informasjon om uttrykksmåten for eit ord, adjektivet *akselbrei*. Her er det ein del spøkefulle seiemå-

tar for å karakterisere folk som er litt for mykje frampå, litt for karslege. Dette har ført til eit underoppslag som ikkje stod i den gamle artikkelen. Vi hadde materiale på det frå før, men via Facebook er desse komne til: «Akselbrei over rauva, å hainnfast te å skjit» (Verdal); «akselbrei over raua og handsterk i låro» (Tingvoll); «akselbrei over rauva og handsterk te å skjit» (Soknedal). «Akselbrei over raua å tong i sessa» (Oppdal) er ikkje redigert inn, men Oppdal er med i heimfestinga. Soknedal er også komme med i heimfestinga på grunn av Facebook. Den publiserte artikkelen kan lesast i Norsk Ordbok (Norsk Ordbok b).

7. Reservasjonar/innvendingar

Med dette har vi sett ved nokre få eksempel at Facebook kan vera med og supplere materialet vårt. Denne kjelda må likevel, som andre, nyttast kritisk.

Postingar på Facebook er, som vi alle veit, ikkje alltid verken ærlege eller pålitelege, og kvaliteten på informasjonen varierer veldig. Trådane utviklar seg ofte i ei skjemtande, spøkefull stemning. Det utartar til rein moro, og tull og tøv. Så ein må vurdere kvar bidragsytar grundig. Når ein arbeider med eit slikt materiale over tid, får ein likevel ein viss nase for kva eller kven som er til å stole på. Ofte blir opplysningane støtta av fleire bidragsytarar, noko som gir ein viss kvalitetskontroll.

Ikkje alt er like interessant heller: Nokså mange av dei orda som blir trekt fram, er i røynda heilt vanlege standardnorske ord, men dei kan bli oppfatta som sjeldne og gammaldagse. Men slike opplysningar er interessante data i seg sjølv, dei har altså ein eigenverdi, og kan indikere ordforråd som kan vera på veg ut av dialekten i framtida.

Eg skal vise eit par små eksempel på opplysningar eller påstandar som nok absolutt er seriøst meint, men som likevel må vera feil. Eksempla er frå ei Facebook-gruppe om dialekten i Sunndal på Nordmøre. Dette er «morsmålet» mitt, så eg kjenner det relativt godt. Vi skal sjå kjapt på eit par ordformer. Begge er jamvektsord, med opphavleg kort rotstaving.

Spesielt for Sunndalen er mykje vokalforlenging i gamle kortstavingar, både ein- og tostava. Herunder kjem altså dei tostava jamvektsorda. Samtidig har sunndalsmålet tradisjonelt runding av gammal lang a til å (*baot*, å *slao*). *Brune* 'eld, varme', uttala *bråne*, har elles i trøndermål former som *brånne*, *brånna* og austleg *brånnå*. På Facebook-gruppa for Sunndal gir

enkelte opp uttalen *braone*. Her må altså nokon ha slutta seg til at det er meir tradisjonsrett å uttala *bråne* med diftong *ao*.

Noko tilsvarande kan vi sjå ved verbet å *kåpa* 'trille'. Dette har ingen brennsikker etymologi (det er ikkje belagt i norrønt), men det må vera kortstava, jf former som *kåppa* og *kåppå* (med konsonantforlenging) i andre trønderdialektar. På sunndaling heiter det å *kåpa* (med jamvektsform, men lang rotvokal), men på Facebook-gruppa hevda somme at uttalen var *kaopa*. Det er likevel umogleg med a-infinitiv etter ein *ao*-diftong, for da må det vera eit opphavleg overvektsord. Det vart da også diskusjon om desse påståtte formene. Eg ser ikkje bort frå at dei som klemte til med *ao*-diftong i *braone* og *kaopa* oppfatta seg som meir genuine dialektbrukarar.

No kan ein kanskje også seie at vi faktisk kan ha ein mekanisme her der jamvektsord med «lydrett» rotvokal å kan få *ao* (gjennom naboopposisjon). Noko liknande har skjedd med eit par former som har festa seg hos ganske mange: *aover* 'over' og *aorntle* 'ordentleg'. Vi kan altså ane ein falsk eller hyperkorrekt *ao*-diftong som tendens. Men no står *ao*-diftongen såpass svakt blant yngre dialektbrukarar at dette neppe vil få særleg stort omfang.

I alle fall viser dette at ein må ha ganske god lokalkunnskap for å tolke opplysningane riktig og ikkje gå i fella.

8. Konklusjon

I prinsippet kan eg ikkje sjå at posteringar på Facebook er meir eller mindre påliteleg enn andre kjelder, som spørjelister, intervju eller lokale ordsamlingar for den del. Ein kjem i kontakt med eit breiare spekter av informantar enn ved tradisjonelle metodar. Eg håpar å ha vist, ved dette vesle glimtet, at det kan vera mykje å hente, i alle fall når det gjeld situasjonen i Norge. Det ser ut som eit nyttig supplement til det vi allereie veit om dialektordforrådet i landet.

Det er også grunn til å merke seg at dei nettverka av informantar vi tidlegare hadde om norske dialektar, ikkje lenger er operative. Informantnettet til Norsk Ordbok, som nemnt tidlegare, er ikkje lenger oppretthalde. Og slett ikkje det som Norsk Målførearkiv i si tid hadde. Dette arkivet er nedlagt for lengst, og materialet er flytta til Bergen, der det høyrer under Språksamlingane. Materialet frå Facebook-gruppene veks stadig, det er

ingen teikn til at dette skal avta. Og det er i ferd med å bli så stort at det er nærmast uoverkommeleg å følgje opp for ein enkelt person. Hittil har ikkje arbeidet heilt vorte systematisert, og det er vanskeleg å seie presist kor mykje tid som er lagt ned på denne delen. Men stort sett kvar einaste dag finn eg noko som kan vera verdt å registrere. Spørsmålet melder seg om det ikkje ville vera nyttig for det leksikografiske miljøet i Bergen å avsette noko ressursar til dette.

Litteratur

Aasen, Ivar 1850. *Ordbog over det norske Folkesprog*. Kristiania: Carl C. Werner.

Aasen, Ivar 1873. *Norsk Ordbog*. Kristiania: Mallings Boghandel.

Aune, Kolbjørn 1976. *Sledenemningar*. Oslo: Norsk Målførearkiv.

Bandle, Oskar 1967. *Studien zur westnordischen Sprachgeographie. Haustierterminologie im Norwegischen, Isländischen und Färöischen. A. textband*. København: Munksgaard.

By, Terje 2022. *Sesån sa vi det. Ord og uttrykk fra Åfjord*. Åfjord: Dialektgruppa, Åfjord Historielag.

Christie, W. F. K. 1937. *Norsk Dialect-Lexicon*. Bergen: Bergen Museum.

Dalen, Arnold 1967. Sele, greie, reiskap. Nemningsbruk i samband med hesteselen. I: *Årbok for Trøndelag*. Trondheim: Trønderlaget, 95–106.

Dalen, Arnold 1985. Nemningar for 'opphovning og smerte i handleden'. I: Bull, Tove & Anton Fjeldstad (red.), *Heidersskrift til Kåre Elstad*. Tromsø: Institutt for språk og litteratur, Universitetet i Tromsø, 208–217.

Donali, Ingeborg & Kari Indseth 2015. *Rørosordboka*. Røros: Røros Museums- og Historielag.

Fjeld, Ruth & Lars Vikør 2008. *Ord og ordbøker. Ei innføring i leksikologi og leksikografi*. Kristiansand: Høyskoleforlaget.

Fjellstad, Joar 2020. *Rennebumålet. Språket i Rennebu ved årtusenskiftet*. Rennebu: Joar Fjellstad.

Hallager, Laurents 1802. *Norsk Ordsamling*. København: Sebastian Popp.

Jenstad, Tor Erik 2001. *Ein repetis i obligadur. Folkemusikkterminologi i norske dialektar; med vekt på feletradisjonen*. Oslo: Novus.

Jenstad, Tor Erik 2011. Deltakande observasjon som leksikografisk metode. I: *Jysk, ømål, riksdansk m.v. Festskrift til Viggo Sørensen & Ove Rasmussen*. Århus: Peter Skautrup-Centeret, 317–323.

Jenstad, Tor Erik 2012. Dialektgrupper på Facebook. I: *Du mitt Nordmøre*. Sunndalsøra: Nordmøre mållag, 76–80.

Norsk Ordbok: ordbok over det norske folkemålet og det nynorske skriftmålet 1966–2016. Oslo: Det norske samlaget.

Norsk Ordbok a. <http://no2014.uib.no/eNo/tekst/tekst_grunnlagsmateriale.html> Henta april 2023.

Norsk Ordbok b. <https://alfa.norskordbok.no/?men=noob&mco=no&mc1=ah&q=akselbrei&but=akselbrei&scope=e> Henta april 2023.

Numedalsmål 2023. <https://www.numedalsmal.no> Henta april 2023.

Opsahl, Toril 2015. Kan ord i bruk bli ord i bok? Urbane ungdomsvarieteter i framtidige ordboksressurser. *LexicoNordica* 22, 131–149.

Ore, Christian-Emil & Åse Wetås 2014. Norsk Ordbok i den digitale tidsalderen. *LexicoNordica* 21, 121–139.

Ross, Hans 1895. *Norsk Ordbog*. Christiania: Cammermeyer.

Urdland Karlsen, Helene, Lars S. Vikør & Åse Wetås (red.) 2016. *Livet er æve, og evig er ordet. Norsk Ordbok 1930–2016*. Oslo: Samlaget.

Vallemål 2019. <https://www.vallemal.no> Henta april 2023.

Vikør, Lars S. & Åse Wetås 2016. Norsk ordbok – om folket – av folket – for folket. I: Urdland Karlsen, Helene, Lars S. Vikør & Åse Wetås (red.), *Livet er æve, og evig er ordet. Norsk Ordbok 1930–2016*. Oslo: Samlaget, 15–39.

Balancen i ordforrådet i en historisk ordbog

Ellert Þór Jóhannsson & Simonetta Battista

This article is concerned with the process of selecting example citations to illustrate forms and meaning of the vocabulary registered in *A Dictionary of Old Norse Prose* (ONP), a historical dictionary project founded in 1939 and based in Copenhagen. After an initial phase of excerption of the relevant text material, the question arose as to whether the dictionary was lacking examples of some common words and expressions. To amend this potential deficit, it was decided early on to excerpt a few key texts extensively, by writing down every single word in context and adding it to the citation archive. In this article we present a case study that describes this endeavor and analyzes the results of this effort to even out a perceived imbalance in the lexical description. The results show that the additional citations unveiled some new definitions as well as phrases. However, by including hundreds of similar examples of common function words, such as conjugations, prepositions and pronouns, there arises a different kind of imbalance the user needs to be aware of.

KEYWORDS: historical dictionary, citations, excerption, dictionary editing

1. Indledning

I denne artikel fokuserer vi på materialet i en historisk ordbog, *Ordbog over det norrøne prosasprog* (ONP). ONP's leksikografiske arbejde med prosatekster fra middelalderen bygger på en samling af citater taget fra videnskabelige udgaver af bevarede håndskrifter efter filologiske principper. Formålet med artiklen er at analysere citatsamlingen i lyset af de leksikografiske udfordringer, man støder på under indsamlingen af materialet. Med udgangspunkt i ONP's selektive excerpering diskuterer vi det, som vi kalder balancen i beskrivelsen af ordforrådet, dvs. hvorvidt og hvor fyldestgørende de excerperede citater repræsenterer det bevarede tekstkorpus.

Efter en indledende fase med excerpering af det relevante tekstmateriale opstod spørgsmålet, om ordbogen manglede eksempler på en række almindelige ord og udtryk. Det førte til en særlig indsats, hvor fem udvalgte tekster blev excerperet ud fra et helhedsprincip, dvs. ord for ord, hvilket resulterede i ca. 38.000 ekstra citater. Vi vil her præsentere en case

study, hvor vi beskriver indsatsen og undersøger, hvordan den har påvirket balancen i det leksikografiske materiale, som i dag findes i ONP.

Artiklen er opbygget på følgende måde: efter denne indledning giver vi i andet afsnit et overblik over ordbogens principper og organiseringen af det leksikografiske materiale. I tredje afsnit beskriver vi den nævnte særlige excerperingsindsats. I fjerde afsnit analyserer vi materialet for at finde ud af, hvor mange og hvilke typer af ord, der blev excerperet fra de fem tekster. Vi sammenligner resultatet med tilgængelige lemmatiserede elektroniske tekster. Vi giver nogle eksempler på, hvordan det særligt excerperede materiale påvirker individuelle ordbogsartikler. Nogle konkluderende bemærkninger følger i femte afsnit.

2. *Ordbog over det norrøne prosasprog*: principper og organisering

Ordbog over det norrøne prosasprog er et ordbogsprojekt ved Københavns Universitet, som registrerer ordforrådet i oldnordiske prosatekster, overleveret i norske og islandske håndskrifter fra ca. 1150 til slutningen af middelalderen. ONP blev grundlagt i 1939 og var oprindeligt tænkt som en slags suppleringsværk til de store ordbøger fra det 19. århundrede, dvs. *Ordbog over det gamle norske sprog* af Fritzner (1886, 1891, 1896) og *An Icelandic-English Dictionary* af Cleasby & Vigfusson (1874). Man indså imidlertid hurtigt, at udarbejdelsen af en mere fyldestgørende beskrivelse af det norsk-islandske middelaldersprog ville kræve et nyt, selvstændigt opslagsværk. ONP blev baseret på en række filologiske principper, der adskilte den fra forgængerne: ordbogen skulle genspejle originalkilderne så nøjagtigt som muligt med unormaliseret ortografi og præcise kildehenvisninger; desuden skulle den omfatte eksempler på alle bevarede ord fra forskellige prosagenrer med "særlig hensyntagen til ældste forekomst af såvel sproglige som syntaktiske fænomener og til kulturhistorisk interessant ordforråd" (Widding 1964:5). Et inspirerende projekt var *Middle-English Dictionary* som har en lignende filologisk tilgang til kildematerialet (cf. MED).

ONP har i tidens løb gennemgået forskellige stadier: et citatarkiv, en delvis trykt udgave (ONP 1–3), og siden 2010, et digitalt leksikografisk værktøj (ONP Online). Udgivelsen af trykte ordbogsbind blev indstillet i 2004, og kort derefter blev en stor del af det udarbejdede materiale gjort tilgængeligt på nettet (cf. Jóhannsson & Battista 2016). Onlineudgaven

er for nyligt blevet fuldstændigt redesignet og bliver løbende udvidet med nye artikler og links til andre relevante ressourcer og digitale hjælpemidler (cf. Jóhannsson, Battista & Wills 2021).

Det grundlæggende excerperingsarbejde var meget nøjagtigt og inkluderede alle værker fra alle bevarede prosagenrer. Resultatet blev et omfattende citatarkiv, der i dag består af over 800.000 repræsentative eksempler fordelt på ca. 65.000 lemmaer. Hele materialet er tilgængeligt og søgbart i ONP Online, dvs. både de færdige artikler med tilhørende citater, og de citater som ikke er redigeret endnu samt citater der af pladsmæssige årsager blev udeladt i de trykte bind. I modsætning til de trykte bind er der ingen pladsbegrænsning i onlineudgaven, hvilket har medført nye redigeringsretningslinjer: alle citater skal redigeres og placeres i strukturerede ordbogsartikler. På nuværende tidspunkt er over halvdelen af citatmaterialet bearbejdet i artikler.

ONP's kildemateriale er fordelt på 437 værker, som hver omfatter et varierende antal tekster. Et værk er defineret som enten samme tekst bevaret i mange forskellige håndskrifter, fx en islændingesaga som *Njáls saga*, der findes i mange håndskrifter i varierende versioner, eller urelaterede tekster, som har et fælles udgangspunkt, fx *Diplomatarium Islandicum,* der består af mange korte, juridiske dokumenter af forskellig slags (se Jóhannsson & Battista 2018 for nærmere diskussion om værker).

Værkerne fordeler sig videre på forskellige genrer. I alt er der ti prosagenrer defineret af ONP: islændingesagaer (fx *Njáls saga*, *Egils saga*); totter: korte fortællinger om islændinge; oldtidssagaer: heltesagn fra den germanske oldtid; samtidssagaer, der omhandler mennesker og begivenheder i forfatternes samtid (fx *Sturlunga saga*); høvisk litteratur: primært riddersagaer, dvs. tekster oversat fra oldfransk og originale værker inspireret af international litteratur; historisk litteratur, primært sagaer om de norske konger; lærde tekster, fx. *Snorra Edda*, men også diverse tekster med videnskabeligt indhold; religiøs litteratur, omfattende fx tekster baseret på latinske, hagiografiske tekster og legendelitteratur oversat fra middelnedertysk og oldengelsk; juridiske tekster, dvs. norske og islandske lovtekster; diplomer: breve og officielle dokumenter, som belyser aspekter af dagliglivet i middelalderen.

I Tabel 1 findes en liste over alle genrer med angivelse af antallet af værker og citater under hver enkel (se også Jóhannsson & Battista 2018). Cirka 1 % af citaterne er ikke klassificeret under en genre, typisk fordi de

i streng forstand falder udenfor ONP's afgrænsede korpus, som fx personnavne eller poetiske ord.

TABEL 1. Fordeling af prosatekster i ONP's citatsamling.

Genrer	**Citater**	**%**	**Værker**	**%**
Religiøse tekster	213773	26,2 %	143	33 %
Historisk litteratur	139495	17,1 %	32	7 %
Islændingesagaer	97227	11,9 %	41	9 %
Høvisk litteratur	82592	10,1 %	52	12 %
Juridiske tekster	77032	9,4 %	50	11 %
Diplomer	61562	7,5 %	4	1 %
Samtidssagaer	58604	7,2 %	11	3 %
Oldtidssagaer	33971	4,2 %	29	7 %
Lærde tekster	24516	3,0 %	10	2 %
Totter	17659	2,2 %	65	15 %
Ukategoriseret	8867	1,1 %		
I alt	**815.298**	**100 %**	**437**	**100 %**

Man kan af Tabel 1 se, at de fleste citater stammer fra religiøse tekster, efterfulgt af historisk litteratur og islændingesagaer. Enkelte andre genrer omfatter også et betydelig antal værker.

3. En særlig excerperingsindsats

Indsamlingen af materialet til ordbogsarbejdet stillede ONP's redaktører over for adskillige udfordringer. Excerpering af tekster var en selektiv proces, hvor redaktørerne skulle bedømme, hvilke citater, der var relevante ifølge ordbogens principper. Det betød samtidig, at interessen rettedes mere mod de mærkelige og usædvanlige former og vendinger, fremfor det almindelige. Efter første fase af excerperingen stod det endnu ikke klart for redaktørerne, hvorvidt det indsamlede materiale repræsenterede ordforrådet i kildeteksterne. Ved en nærmere gennemgang af citatsamlingen viste der sig at være en vis ubalance, idet man havde mange eksempler på mærkelige ord og former, men ikke særligt mange eksempler på ganske

almindelige ord, funktionsord, så som præpositioner, konjunktioner og pronominer, selvom disse ord forekommer hyppigt i tekstmaterialet. Ordbogens redaktion vurderede, at man formentlig gik glip af vigtige eksempler på sprogbrug og almindelige vendinger, som hørte til en udførlig leksikalsk beskrivelse af sproget. For at rette op på denne ubalance igangsatte man et særligt excerperingsprojekt, der skulle registrere hvert eneste ord fra nogle udvalgte, repræsentative, ældre tekster.

Dette blev muliggjort takket være eksterne ressourcer, som omtales i en tidlig rapport om ordbogsarbejdet.

> På ganske uventet måde fik ordbogen øget arbejdskraft under krigen. Ved den såkaldte nødbeskæftigelse blev der af Socialministeriet stillet beløb til rådighed til at beskæftige arbejdsledige [...] De anviste medhjælpere kunne uden forkundskaber udskrive visse karakteristiske og centrale tekster efter thesaurusprincippet: hvert ord med sit citat på sin seddel [...] Alle disse sedler blev mærket på særlig måde (stemplet med et A i øverste venstre hjørne) […] Særlig med hensyn til formord [dvs. funktionsord] eller småord har denne form for seddeludskrivning haft sin store betydning. (Widding 1964:7)

Fem tekster blev udvalgt ud fra følgende kriterier: der var tale om originale tekster, overleveret i forholdsvis gamle håndskrifter og tilgængelige i videnskabelige udgaver. Teksterne var i øvrigt særligt relevante fra et kulturhistorisk synspunkt. Det drejede sig om følgende tekster:

- *Gylfaginning* fra *Snorra Edda* (GKS 2367 4°, også kaldet *Konungsbók Snorra Eddu*, ca. 1300-1350): *Snorra Edda* er et enestående værk i nordisk middelalderlitteratur. Det er skrevet af Snorri Sturluson (1179–1241), en prominent forfatter og høvding i Island. *Gylfaginning* er en vigtig kilde til nordisk mytologi, idet størstedelen af vores viden om de nordiske guder er bevaret her.

- *Heiðarvíga saga* (Holm perg 18 4°, ca. 1300): en kendt islændingesaga, som er kun delvis bevaret i dette håndskrift. Håndskriftet blev tidligt delt op, og den del af det, som inkluderede første del af sagaen, gik tabt ved Københavns brand i 1728. Den anden del befandt sig i Sverige og overlevede derfor branden. Kun den del af sagaen, som er bevaret i den overlevede del af håndskriftet, er excerperet.

- et fragment af *Egils saga Skallagrímssonar* (AM 162 A θ fol, ca. 1250): det ældste bevarede fragment af en af de mest berømte islændingesagaer.

- *Íslendingabók* (AM 113 folx, 1651): et enestående værk om Islands historie, forfattet af Ari Þorgilsson i 1100-tallet, men bevaret i et forholdsvis ungt papirhåndskrift, som imidlertid betragtes som en meget nøjagtig afskrift af en nu tabt skindbog, der formentlig skal dateres til ca. 1200.

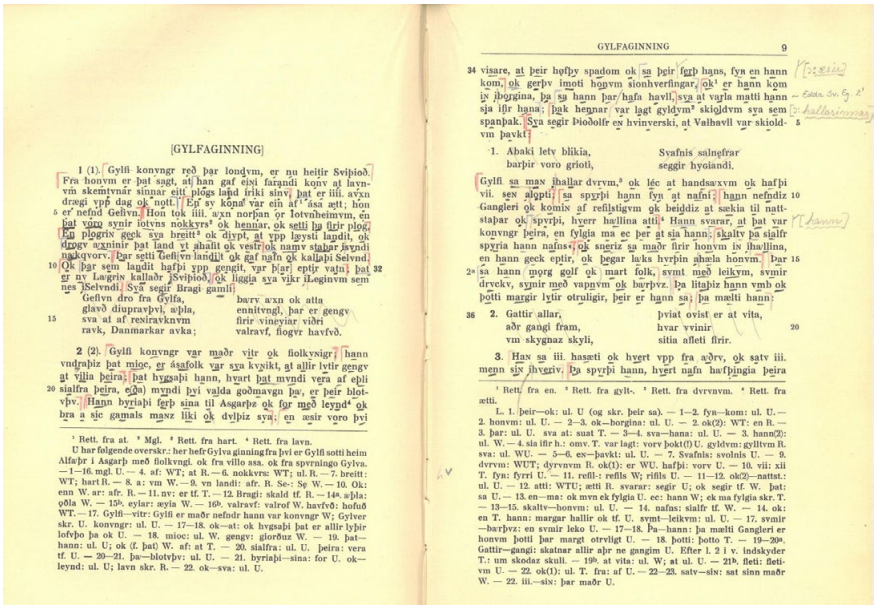
- *Óláfs saga Tryggvasonar* (Holm perg 18 4°, ca. 1300): et vigtigt værk i islandsk litteraturhistorie, det første værk, der fokuserer på en enkelt konges historie. Sagaen findes i to versioner. Versionen i Holm perg 18 er forfattet af munken Oddr Snorrason i 1100-tallet. Denne tekst er bevaret i samme håndskrift som *Heiðarvíga saga*.

De fem tekster repræsenterer kun tre genrer, islændingesagaer (*Heiðarvíga saga* og *Egils saga*), historisk litteratur (*Íslendingabók* og *Óláfs saga*) og lærde tekster (*Gylfaginning*) (jf. Tabel 1). Det drejer sig om centrale genrer, der repræsenterer en betydelig del af det bevarede tekstkorpus. Samtidig bemærker man, at de udvalgte tekster ikke inkluderer oversættelser, juridisk sprog samt tekster fra genrer, som primært er forfattet i det 13. århundrede, fx høvisk litteratur og samtidssagaer.

I lyset af den senere udvikling i norrøn litteraturhistorie kan man måske argumentere for, at disse fem tekster blev fremhævet som de mest repræsentative for det norrøne ordforråd i en periode, hvor vægten blev lagt på bestemte genrer og karakteristikker. Teksterne repræsenterer forholdsvis velkendte og vigtige værker, der blev forfattet i Island allerede i 1100- og 1200-tallet. Måske har man betragtet dem som mere oprindelige eller tættere på højdepunktet i norrøn prosalitteratur end andre, og derfor som en slags guldstandard for sproget. En anden vigtig grund kan være, at udover at være relevante, var teksterne tilgængelige i pålidelige udgaver og forholdsvis nemme at transskribere for "medhjælpere … uden forkundskaber", som det blev beskrevet i rapporten (Widding 1964:7).

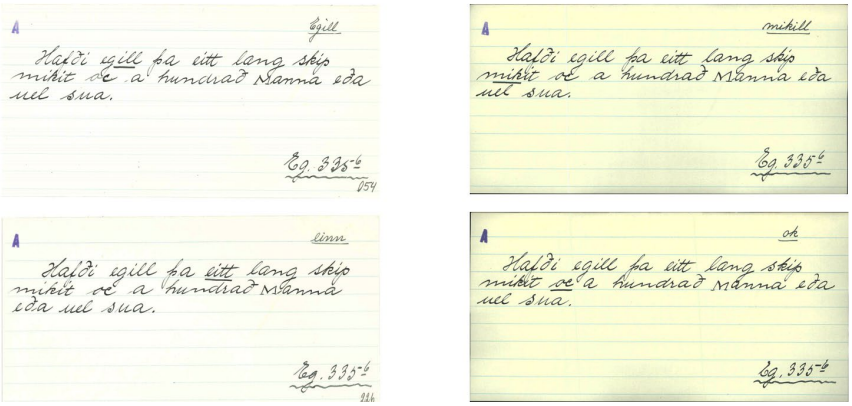
Excerperingen gik ud på at give eksempler på hvert eneste ord i sammenhæng. Tekstudgaverne blev gennemlæst af en redaktør, og hvert ord blev markeret som opslagsord (med understregning) i et citat (indesluttet

i hager) som skulle transskriberes på en seddel. Figur 1 viser et opslag fra den fuldt excerperede udgave af *Gylfaginning*.



FIGUR 1. Opslag fra en fuldstændig excerperet udgave, hvor alle ord er markeret som opslagsord (sort prik under ord) og citater er afgrænset (røde hager).

Resultatet af excerperingsarbejdet var en række sedler, ofte skrevet med samme hånd, hvor hvert citat forekommer under hvert eneste opslagsord, cf. Figur 2.



FIGUR 2. Fire sedler, skrevet med samme hånd, som viser samme citat under fire forskellige opslagsord.

Excerperingsindsatsen resulterede i tilføjelsen af 37.880 nye citater til ONP's citatsamling. Tallene kommer fra ONP's gamle arbejdsnoter, hvor man har registreret de forskellige faser af seddelopskrivningen. I næste afsnit diskuterer vi, hvilken indflydelse denne excerperingsindsats havde på balancen i citatmaterialet og hvordan den figurerer i det leksikografiske materiale, som i dag er tilgængeligt i onlineudgaven af ONP.

4. Undersøgelse og resultat

For at finde ud af hvordan excerperingsindsatsen har påvirket citatsamlingen har vi undersøgt de citater fra hver af de fem tekster som findes i dag i ONP's database. Resultatet vises i tabel 2. Det angives for hver tekst, hvor stor en andel de særligt excerperede citater udgør af det samlede antal citater i ONP.

TABEL 2. Særligt excerperede citater fra hver af de fem tekster sammenlignet med, hvor mange citater der findes i alt i ONP's database fra den pågældende tekst.

Tekst	**Genre**	**Særlige citater**	**%**	**I alt i ONP**	**% af genre**
Íslendingabók	Hist	3996	102 %	3899	2,8 %
Óláfs saga	Hist	4814	76 %	6279	4,5 %
Snorra Edda	Lær	15508	84 %	18483	75,0 %
Egils saga frg.	Isl	2986	29 %	10447	10,7 %
Heiðarvíga saga	Isl	10576	79 %	13439	13,8 %
I alt		**37880**		**52547**	

Tabel 2 viser, hvor mange citater, der blev excerperet fra hver af de fem udvalgte tekster i forhold til det samlede antal citater for hele den pågældende tekst. De fleste tekster var allerede blevet excerperet på normal vis. Man ville derfor forvente, at det samlede antal citater skulle være højere end antallet af citater fra de udvalgte tekster. To tal er dog lidt anderledes end forventet. Det lave tal for *Egils saga* skyldes, at kun et fragment af sagaen blev excerperet ord for ord. Tallene for *Íslendingabók* (102 %) tyder på, at den ikke har været excerperet i forvejen, og at der har været nogle dubletter, som så er blevet slettet under redigeringsprocessen. De 52.547 citater fra disse tekster svarer til 6,4 % af alle citaterne i ONP's database. De 16 hyppigste ord fra hver tekst vises i tabel 3.

TABEL 3. Tabellen viser de 16 hyppigste ord fra hver af de fem tekster med angivelse af ord, ordklasse og antal citater.

Snorra-Edda			Heiðarvíga saga			Egils saga			Íslendingabók			Óláfs saga		
Ord	**Kl.**	**Cit.**	**Ord**	**Kl.**	**Cit.**	**Ord**	**Kl.**	**Cit.**	**Ord**	**Kl.**	**Cit.**	**Ord**	**Kl.**	**Cit.**
ok	*conj.*	966	ok	*conj.*	674	ok	*conj.*	129	vera	*vb.*	172	ok	*conj.*	402
vera	*vb.*	663	sá	*pron.*	389	vera	*vb.*	101	ok	*conj.*	168	sá	*pron.*	208
hann	*pron.*	596	vera	*vb.*	383	sá	*pron.*	87	sá	*pron.*	140	hann	*pron.*	196
sá	*pron.*	511	hann	*pron.*	323	hann	*pron.*	75	er	*conj.*	120	vera	*vb.*	167
er	*conj.*	429	nú	*adv.*	187	hafa	*vb.*	61	en	*conj.*	109	er	*conj.*	104
þá	*adv.*	410	er	*conj.*	161	er	*conj.*	59	hann	*pron.*	105	í	*præp.*	81
en	*conj.*	311	til	*præp.*	160	en	*conj.*	57	þat	*pron.*	83	þá	*adv.*	73
í	*præp.*	234	hafa	*vb.*	131	til	*præp.*	54	þá	*adv.*	77	til	*præp.*	71
þat	*pron.*	223	þat	*pron.*	130	þú	*pron.*	51	í	*præp.*	73	koma	*vb.*	68
svá	*adv.*	217	barði	*sb.*	123	þat	*pron.*	50	inn	*adv.*	67	með	*præp.*	64
til	*præp.*	184	ek	*pron.*	122	þá	*adv.*	49	sonr	*sb.*	62	munu	*vb.*	59
segja	*vb.*	182	í	*præp.*	120	þar	*adv.*	40	hafa	*vb.*	60	ek	*pron.*	59
hafa	*vb.*	156	þú	*pron.*	117	svá	*adv.*	33	maðr	*sb.*	55	þat	*pron.*	58
heita	*vb.*	155	maðr	*sb.*	106	segja	*vb.*	33	á	*præp.*	53	þar	*adv.*	57
hon	*pron.*	145	koma	*vb.*	105	mál	*sb.*	33	til	*præp.*	48	maðr	*sb.*	53
allr	*adj.*	136	þar	*adv.*	101	í	*præp.*	31	vetr	*sb.*	37	sinn	*pron.*	51

De fleste ord er funktionsord af forskellige slags: primært konjunktioner, såsom *ok* 'og', *er* 'som' og *en* 'og, men', men også pronominer som *hann* 'han' og *sá* 'denne' og præpositioner. Der findes også andre ordklasser,

gængse verber som *vera* 'være', *segja* 'sige', *koma* 'komme', substantiver som *maðr* 'mand', og adverbier som *þá* 'da' og *nú* 'nu'.

En betydelig del af citaterne i ONP's materiale, der illustrerer funktionsordene og andre almindelige ord, er excerperet fra de fem udvalgte tekster. Tabel 4 viser de 16 hyppigste opslagsord fra de fem tekster, og hvor stor en procentdel citaterne derfra udgør af det samlede antal citater i ONP for det pågældende lemma.

TABEL 4. Opslagsord, antal citater og % af særlige citater.

Ord	Klasse	Fra de 5	ONP i alt	%
ok	*conj.*	2339	3230	72 %
vera	*vb.*	1486	3240	46 %
sá	*pron.*	1335	2298	58 %
hann	*pron.*	1295	1694	76 %
er	*conj.*	873	3087	28 %
þá	*adv.*	703	1291	54 %
en	*conj.*	579	796	73 %
þat	*pron.*	544	1053	52 %
í	*præp.*	539	1288	42 %
til	*præp.*	517	1669	31 %
hafa	*vb.*	454	2475	18 %
ek	*pron.*	368	745	49 %
maðr	*sb.*	344	1325	26 %
koma	*vb.*	326	3340	10 %
þú	*pron.*	308	444	69 %
nú	*adv.*	294	684	42 %

Mellem halvdelen og tre fjerdedele af eksemplerne, som indeholder hyppigt brugte konjunktioner og pronominer, kommer fra disse tekster, fx *ok* = 72 %, *hann* = 76 %, *en* = 73 %. Samtidig kan man se, at den selektive excerpering af mere betydningsfulde ord alligevel har været "fornuftig". Fx har *maðr* proportionelt flere citater fra andre værker end de fem udvalgte. Det samme kan man sige om verber som *hafa, koma* og *vera*, som har mange betydningsnuancer.

For at sætte resultatet fra undersøgelsen ind i videre sammenhæng kan man sammenligne det med moderne ressourcer. Der findes en del elek-

troniske tekstudgaver, hvor enkelte norrøne tekster er blevet fuldstændigt transskriberet fra et enkelt håndskrift og lemmatiseret. Mange af disse tekster er tilgængelige online i *Arkiv for nordiske middelaldertekster* (MENOTA), der samler elektroniske udgaver af norrøne tekster og giver adgang til dem og relateret data. Resultatet af vores undersøgelse kan sammenlignes med et antal eksempler fra relevante tekster i MENOTAs arkiv. Tabel 5 viser en oversigt over de mest almindelige ord i de norrøne tekster som findes i MENOTA.

TABEL 5. De 16 hyppigst forekommende ord i elektroniske udgaver af norrøne tekster (MENOTA).

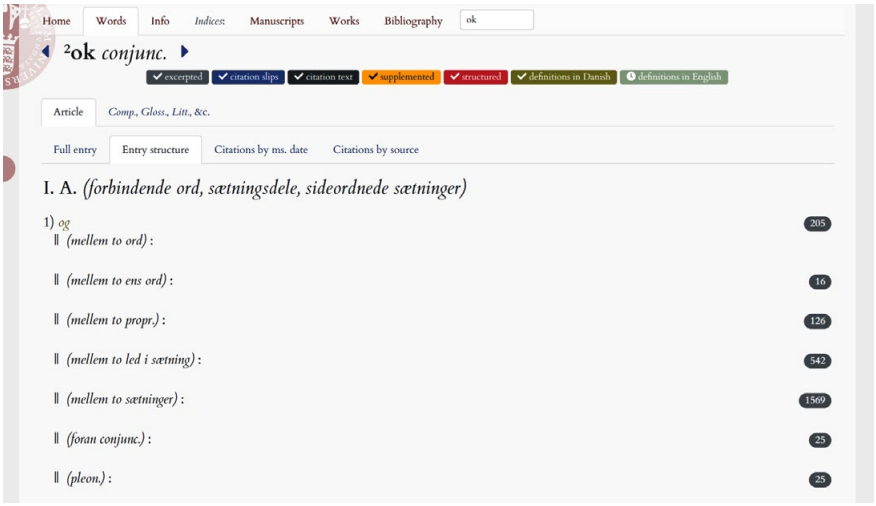
Ord	**Ordklasse**	**Antal**
ok	*conj.*	26154
hann	*pron.*	21810
sá	*pron.*	21757
vera	*vb.*	13672
er	*conj.*	11135
at	*conj.*	8930
í	*prep.*	7517
þá	*adv.*	6729
til	*prep.*	6485
hafa	*vb.*	5809
með	*prep.*	5573
sem	*conj.*	5066
ek	*pron.*	5028
sinn	*pron.*	4858
svá	*adv.*	4810
allr	*adj.*	4803

Når man sammenligner tabel 5 med tabel 4, er det stort set de samme ord som figurerer begge steder, fx konjunktionen *ok*, verbet *vera* og pronomenet *hann*. Den største forskel er, at man i ONP har pronomenet *þat* 'det', som ikke forekommer som selvstændigt lemma i MENOTA-teksterne.

Grunden hertil er, at mange tekstudgivere har valgt at lemmatisere *þat* som en bøjningsform under lemmaet *sá*, som *þat* i mange tilfælde er neutrum form af.

Man lægger også mærke til, at der i MENOTA-teksterne forekommer nogle ord med mange eksempler, som ikke er lige så godt dokumenteret i ONP's data. Det er især tilfældet ved konjunktionen *at*, for hvilken man ikke finder så mange excerperede citater, selvom man ville have forventet, at ordet optrådte hyppigt i de fem udvalgte tekster. Når man ser nærmere på dataene, viser det sig, at konjunktionen *at* er blevet skrevet op på sedler på samme måde som alle andre ord fra de fem tekster. Uoverensstemmelsen skyldes at *at* blev redigeret i det første bind af den trykte udgave (ONP 1). Den gang var praksis ved redigeringsarbejdet, at alle citatsedlerne blev indtastet i ordbogens database, men kun et udvalg blev brugt ved den endelige redaktion af opslagsord med mange eksempler. Ved konjunktionen *at* har man afveget fra denne fremgangsmåde, idet man tidligt i redigeringsprocessen indså, at der var disproportionalt mange eksempler på ordet. Derfor blev ikke alle citater tastet ind, men derimod kun et udvalg. Samme fremgangsmåde blev brugt for andre funktionsord tidligt i alfabetet, fx præpositionerne *af* 'af' og *á* 'på'.

Alle citater, som hører under opslagsord, der er blevet redigeret efter ONP 1–3 udkom, er til gengæld blevet en del af ONP's database, og de er blevet redigeret eller skal redigeres i de færdige artikler. Til sammenligning kan man se, at konjunktionen *ok* ikke er blevet reduceret på samme måde som *at*, idet den har 3230 eksempler, som alle sammen indgår i en struktureret ordbogsartikel. Figur 3 illustrerer delvis artiklens struktur. Det er angivet, hvor mange citater, der hører til hver betydning.



FIGUR 3. Første del af artiklen *ok* i ONP, hvor der kun vises struktur og antal citater.

På Figur 3 kan man se, at der findes 1569 citater, som viser brugen af *ok* "mellem to led i sætning". Man må konstatere, at den fuldstændige excerpering af tekster og inklusion af alle citater i ordbogens database også kan give et skævt billede af ordforrådet med en overvægt af ensartede eksempler.

Der er dog nogle tilfælde, hvor den særlige excerperingsindsats resulterede i nye betydninger og fraser. Figur 4 viser eksempler fra artiklen *mikill* adj. 'megen, stor', hvor den særlige excerpering afslørede både en ny betydning og nogle fraser, som ikke var registreret tidligere. Fra *Óláfs saga* har man et eksempel, hvor adjektivet refererer til stærkt lys og fra *Heiðarvíga saga* nogle vendinger som *ríða mikinn dyn* 'ride med meget støj' og *gera mikit af sér* 'gøre en stor indsats på egen hånd', der illustrerer dette.

FIGUR 4. Udvalgte dele af artiklen *mikill* adj. 'megen, stor' fra ONP.

Disse eksempler viser, at den særlige excerperingsindsats kunne resultere i mere nøjagtig leksikalsk beskrivelse. Dette er dog ikke tilfældet ved de fleste funktionsord.

5. Konklusion

I denne artikel har vi fokuseret på ONP's citatsamling i lyset af de leksikografiske udfordringer, man støder på under indsamlingen af materiale og diskuteret ordforrådet fra de fem særligt udvalgte tekster og hvordan de figurerer proportionelt i materialet og i individuelle ordbogsartikler. Formålet med den særlige excerperingsindsats var at rette op på en formodet ubalance i citatsamlingen med henblik på en mere fyldestgørende leksikalsk beskrivelse. Denne case study har vist at den særlige excerperingsindsats indbragte flere eksempler på underrepræsenterede ord, fx konjunktioner, pronominer, og gængse verber. I flere tilfælde resulterede det dog i mange unødvendige eksempler på samme slags sprogbrug og samme betydning af helt almindelige ord. Man har også kunnet dokumentere nye betydninger og vendinger, omend ikke så mange, som man måske havde håbet på, da man gik i gang med at supplere citatsamlingen på denne måde. Med forsøget på at rette op på den antagne ubalance opstod der en anden slags skævhed i form af et overvældende antal ensartede citater, som ordbogsbrugerne skal være opmærksomme på.

ONP's forsøg med fuld excerpering af de fem tekster kunne fra et moderne synspunkt betragtes som en slags korpuslingvistik, dvs. at inkorporere alle ord fra udvalgte tekster i ordbogen. Det er imidlertid uhensigtsmæssigt at anvende denne metode, hvis man fokuserer på den leksikalske beskrivelse af sproget, fordi datamængden bliver uoverskuelig (som vi har set i forbindelse med artiklen *ok*). Materialet skal afgrænses med henblik på overskuelige, brugervenlige ordbogsartikler, som samtidig skal dække alle nuancerne i sproget. ONP Online viser alle citater, som findes i citatsamlingen, men dette princip skal muligvis revideres i forbindelse med beskrivelsen af funktionsord og andre almindelige ord som kan have hundredvis af eksempler.

Inkorporeringen af korpusmateriale er dog muligt med sprogteknologiske værktøjer og kan med fordel bruges til at supplere ordbogsdata. Man kan benytte elektroniske udgaver af tekster og andre digitale ressourcer til at give brugerne mere fyldestgørende oplysninger om ordforrådet også

til andre formål end at slå en betydning op, fx overblik over mulige grammatiske former, forekomst i andre ordbøger, forskellige tekstudgaver osv. Tilknytning til eksterne ressourcer er noget som ONP arbejder på sideløbende med redigeringsarbejdet, og det er blevet diskuteret andre steder (fx Wills, Jóhannsson & Battista 2018, 2021 og Wills & Jóhannsson 2019). Supplerende citater kan godt stilles til rådighed for brugerne, men et højt antal eksempler kan virke forstyrrende på artikelstrukturen.

Litteratur

Cleasby, Richard & Guðbrandur Vigfússon. 1874. *An Icelandic-English Dictionary*. Oxford: Clarendon Press.

Fritzner, Johan 1886, 1891, 1896. *Ordbog over Det gamle norske Sprog 1–3*, rev. udg. Kristiania: Den norske forlagsforening.

Jóhannsson, Ellert Þór & Simonetta Battista 2018. Middelaldertekster som sproglig ressource. I: Ásta Svavarsdóttir & Helga Hilmisdóttir (red.) *Nordiske Studier i Leksikografi 14: Rapport fra 14. Konference om Leksikografi i Norden, Reykjavík 30. maj–2. juni 2017*, 152–161.

Jóhannsson, Ellert Þór & Simonetta Battista 2016. Ordbog over det norrøne prosasprog Online: struktur og brug. I: Gudiksen, Asgerd & Henrik Hovmark (red.) *Nordiske Studier i Leksikografi 13: Rapport fra 13. Konference om Leksikografi i Norden, København 19.–22. maj 2015*, 165–175.

Jóhannsson, Ellert Þór, Simonetta Battista og Tarrin Wills 2021. Legacy data in a digital age. I: Reinsone, Sanita, Inguna Skadiņa, Jānis Daugavietis & Anda Baklāne (red.) *Digital Humanities in the Nordic Countries 2020, Post-Proceedings of the 5th Conference Digital Humanities in the Nordic Countries Riga, Latvia, October 21–23, 2020*, 248–254.

MED = *Middle English Dictionary*. Ed. Robert E. Lewis, et al. Ann Arbor: University of Michigan Press, 1952-2001. Online edition in Middle English Compendium. Ed. Frances McSparran, et al. Ann Arbor: University of Michigan Library, 2000–2018. <http://quod.lib.umich.edu/m/middle-english-dictionary/>. Hentet marts 2023.

MENOTA = *Medieval Nordic Text Archive. Arkiv for nordiske middelaldertekster*. <https://menota.org>. Hentet august 2022.

ONP 1–3 = Degnbol, Helle, Bent C. Jacobsen, James E. Knirk, Eva

Rode, Christopher Sanders & Þorbjörg Helgadóttir (red.) *Ordbog over det norrøne prosasprog / A Dictionary of Old Norse Prose. ONP Registre* 1989. *ONP 1: a–bam* .1994. *ONP 2: ban–da.* 2000. *ONP 3: de–em.* 2004. *ONP Nøgle/Keys.* 2004. København: Den Arnamagnæanske Kommission.

ONP Online = *Ordbog over det norrøne prosasprog / A Dictionary of Old Norse Prose*. <http://onp.ku.dk>. Hentet august 2022.

Widding, Ole 1964. *Den Arnamagnæanske Kommissions Ordbog, 1939–1964: Rapport og plan*, København: G.E.C.GADS Forlag.

Wills, Tarrin & Ellert Þór Jóhannsson 2019. Reengineering an Online Historical Dictionary for Readers of Specific Texts. I: Kosem, Istok, (red.) *Electronic lexicography in the 21st century: Smart lexicography: Proceedings of the eLex 2019 conference*. Brno: Lexical Computing CZ, 116–129.

Wills, Tarrin, Ellert Þór Jóhannsson & Simonetta Battista 2018. Linking Corpus Data to an Excerptbased Historical Dictionary. I: Čibej, Jaka, Vojko Gorjanc, Istok Kosem & Simon Krek (red.) *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*. Ljubljana: Ljubljana University Press, Faculty of Arts, 979–987.

Wills, Tarrin, Ellert Þór Jóhannsson & Simonetta Battista 2021. Integrating TEI/XML Text with Semantic Lexicographic Data. I: Reinsone, Sanita, Inguna Skadiņa, Jānis Daugavietis & Anda Baklāne (red.) *Digital Humanities in the Nordic Countries 2020, Post-Proceedings of the 5th Conference Digital Humanities in the Nordic Countries Riga, Latvia, October 21–23*, 2020, 16–25.

To islandske ordbøgers ordforråd i hundrede år

Halldóra Jónsdóttir & Þórdís Úlfarsdóttir

Two Icelandic dictionaries were published one hundred years apart: *Islandsk-dansk ordbog* (1920-1924) by Sigfús Blöndal and *Ordbog over moderne islandsk* (2021) (*Íslensk nútímamálsorðabók*), which is web-based from the start. The former and considerably larger work contains 108,000 lemmas whereas the latter has 56,000 lemmas. This study briefly outlines the origin of the vocabulary of the two dictionaries. A small selection of compounds is in focus: compounds having the first part *brauð-* ('bread-') and the last part *-fólk* ('-people, -folk'), as these two words belong to the central vocabulary, and a comparison is made between these compounds. Moreover, a computerised comparison process reveals that the number of entries common for both dictionaries is 24,000 lemmas. These may tentatively be regarded as the core vocabulary of Icelandic in the 100 year period.

KEYWORDS: Icelandic, historical dictionary, vocabulary

1. Indledning

Artiklen beskriver en undersøgelse af ordforrådet i to islandske ordbøger der udkom med hundrede års mellemrum: *Islandsk-dansk ordbog* (*Íslensk-dönsk orðabók*) (1920-24) af Sigfús Blöndal (herefter Blöndals ordbog) og *Ordbog over moderne islandsk* (*Íslensk nútímamálsorðabók*) (2021), redigeret af Halldóra Jónsdóttir og Þórdís Úlfarsdóttir (herefter OMI). Den førstnævnte udkom på tryk, mens den sidstnævnte kun findes på webben. Efter at den digitaliserede udgave af Blöndals ordbog udkom i 2021, blev der åbnet op for muligheder for at granske ordbogens indhold i større omfang end tidligere, blandt andet at sammenligne ordforrådet i Blöndals ordbog med andre kilders. I artiklen vil vi give en karakteristik af hver af de to ordbøger og foretage en sammenligning af deres indhold. Undersøgelsen omfatter ændringer i ordforrådet over en periode af hundrede år, eksemplificeret ved udvalgte sammensatte ord. Med udgangspunkt i disse to ordbøger spørger vi endvidere om det eventuelt er muligt at definere kerneordforrådet i islandsk. Det skal dog bemærkes at undersøgelsen er begrænset, og at vores gennemgang kun omfatter en lille del af ordbøgernes indhold.

2. Ordforrådets omfang og oprindelsebøger

Blöndals ordbog udkom på tryk i to bind 1920-1924. Ordbogens kildesprog er islandsk med danske ordforklaringer, men den har mange fællestræk med en monolingval ordbog. Det danske målsprog har hovedsagelig en politisk forklaring, nemlig at Island var en del af Danmark indtil 1918, og de to landes statskasser bekostede værket i fællesskab. Blöndals ordbog indeholder 108.000 lemmaer, og er hidtil det største ordbogsværk som beskriver det islandske sprog. I 1963 blev der udgivet et supplementsbind som indeholder knap 40.000 lemmaer, men dette ordforråd indgår ikke i vores undersøgelse.

OMI, den anden ordbog i vores sammenligning, er en monolingval webordbog der blev åbnet i 2016 (selvom den stadig er og var under redigering 2021 da undersøgelsen blev gennemført). Den indeholder 56.000 lemmaer og er altså betydeligt mindre end det ældre værk. OMI er den nyeste monoligvale ordbog over islandsk og er fra starten designet specielt for webben. Den bygger på nyt materiale, herunder korpusser (se kapitel 2.2), og er udarbejdet med nye metoder.

Disse to islandske ordbøger, som udkom i hver sit århundrede, er blevet undersøgt og sammenlignet med henblik på ordforrådet. Projektet kunne selvfølgelig blive meget omfangsrigt, og det er ikke muligt at give en dyb analyse i en kort artikel.Vi valgte derfor at inddrage et begrænset udvalg ord i vores sammenligning.

2.1. Blöndals ordbog

Sigfús Blöndal oplyser om sine kilder i det grundige forord til sin ordbog. Han brugte sammen med sin hustru, Björg Þ. Blöndal, adskillige år på at excerpere forskellige islandske skrifter som skulle danne grundlaget for ordbogen. Der blev inkluderet mange eksisterende ordsamlinger i ordbogen, og redaktørerne havde desuden et stort antal informanter i hele Island som dokumenterede lokalsprog og dialekter (jf. Jónsdóttir & Úlfarsdóttir 2021:162). Blöndal lagde stor vægt på at repræsentere almuesprog og talesprog, f.eks. anføres ”gadedrenge i Reykjavík” som en kilde. Blöndal omtaler dette i sin fortale:

> Det Utal af Nydannelser og af ny Betydninger i gamle Ord, som en voksende og stadig mere mangeartet Kulturudvikling havde medført, maatte der tages tilbørligt Hensyn til. Og det uforfalskede Almuesprog, som kun delvis fremtræder i vor Litteratur, var noget af det vigtigste. (Fortale: VII).

Lemmaerne i Blöndals ordbog tager udgangspunkt i den tids sprog og indeholder det vigtigste islandske ordforråd omkring århundredeskiftet 1900. Ordbogen giver en nøjagtig analyse af det centrale ordforråd, og nogle af ordbogsartiklerne er meget indholdsrige. Denne undersøgelse har påvist at ordbogen desuden medtager et stort antal *ad hoc* sammensætninger, ord som man nu anser for at "sige sig selv", f.eks. *aðaltröll* ('hovedtrold'), *aðalvísindi* ('hovedvidenskab'), *hálfdimma* ('halvmørke') og *hálfdimmur* ('halvmørk').

I ordbogen findes der endvidere et antal lemmaer som er konstruerede ord. Grunden til deres tilstedeværelse er uden tvivl at danne modpart til almindelige danske begreber: *aðgöngumiðaokrari* 'billetsjover' og *gróðurneyslumaður* 'vegetar' (jf. Bjarnadóttir 1997).

Under ordbogens tilblivelse i de to første årtier af det tyvende århundrede var dansk det vigtiste fremmedsprog i Island, og i mange tilfælde nøglen til andre fremmedsprog. Ordbogens redaktion var optaget af at finde frem til nye islandske ord for alle slags fænomener, bl.a. ting som kun fandtes i udlandet på den tid. Det er grunden til det store antal neologismer og låneord, hvoraf mange ikke havde fået fodfæste i islandsk, men som redaktionen har anset for vigtigt at samle på ét sted.

Ordsamlinger fra enkeltpersoner, trykte såvel som utrykte, blev som sagt også inkluderet i ordbogens lemmaliste, foruden at den indeholder et stort antal poetiske og historiske ord. I sin fortale diskuterer Blöndal denne del af sin ordselektion:

> Jeg har i det hele taget med særlig Forkærlighed søgt at belyse alt hvad der angik Islands folkelige Kultur i videste Forstand: Overtro, Skikke, Redskaber, Arbejdsmetoder o.s.v. Nu da den ny Tid vælter ind over Landet, har jeg efter ævne villet bidrage til at redde en Del fra glemsel medens der endnu er Tid dertil. (Fortale: XII)

Det bemærkes her at det danske ordforråd i Blöndals ordbog ikke hidtil har fået stor opmærksomhed, og det kalder på nærmere studier, specielt

med henblik på at det hovedsagelig var islændinge som skrev ordforklaringerne. Dette kræver en anden undersøgelse, som man kan forestille sig blive interessant.

2.2. Ordbog over moderne islandsk

OMI er udgivet af Árni Magnússon instituttet for islandske studier og er baseret på den islandske del af den nordiske ordbog ISLEX (jf. Úlfarsdóttir 2013; Jónsdóttir & Úlfarsdóttir 2019 og 2020). ISLEX stammer oprindelig fra en stor ordbogsbase som blev udarbejdet i nittenhundrede og halvfemserne som et nordisk initiativ. Det tog seks år at udarbejde ordbogen med seks nordiske målsprog, men det materiale der opstod under ordbogsarbejdet, har siden dannet grundlaget for et antal ordbogsprojekter.

Det er klart at metoderne for ordbogsarbejde er helt andre i dag end for et hundrede år siden. Et af Blöndals formål med sin ordbog var at medtage et alsidigt ordforråd: "den skal søge at give et uforfalsked og sanddru Billede af Sproget som det er, og belyse dets Kringelkroge saavidt muligt" (Sigfús Blöndal 1920-24:XII). Nu er metoderne andre og der skal være en anledning til en leksikografisk analyse af de valgte lemmaer, semantisk og/eller syntaktisk.

Før i tiden var hele arbejdsprocessen som bekendt manuel og meget tidskrævende, men nu er der god adgang til omfangsrigt kildemateriale. Desuden har kriterierne for indvælgelse af ordforråd i ordbøger ændret sig betydeligt. Efter 2000 er der sket store fremskridt inden for sprogteknologi som affødte nye sproglige resurser, og der blev opbygget nogle islandske tekstkorpusser som har spillet en afgørende rolle for ordbogsarbejdet. Foruden at fremskaffe nyt materiale kan korpusser anvendes til forskellige andre formål, f.eks. kan man se ordene i tekstsammenhæng samt ordets alder, betydning og anvendelsesområde. Ved redaktionen af OMI er der blandt andet blevet anvendt frekvenslister fra *Risamálheildin* ('Gigakorpus') (2018), som er et nyt, stort islandsk korpus (Steingrímsson et al. 2018). I årene 2019-2021 blev der tilføjet 5.700 lemmaer til OMI som hovedsagelig stammede fra frekvenslister fra dette korpus. Frekvenslisterne repræsenterer et nyt ordforråd som bliver kandidater til nye opslagord. Alligevel siger frekvenslisterne ikke alt eftersom de indeholder mange *ad hoc* sammensætninger og andet som ikke nødvendigvis skal med i ordbogen efter redaktørernes vurdering. Lemmalisten er ligeledes blevet udvidet med nyt ordforråd fra flere kilder, f.eks. brugerhenvendelser og redaktørernes egne iagttagelser.

OMI er under regelmæssig opdatering eftersom sproget er i stadig fornyelse, nye fænomener opstår og gamle forsvinder, og ordforrådet ændres i takt med dette. Til visse områder som f.eks. computere, teknik og medicin, miljø og samfund tilføjes der især mange ord (se en nærmere beskrivelse af indvælgelse af ord i Jónsdóttir & Úlfarsdóttir 2019:14-17).

3. Sammenligning af ordforrådet

3.1. Kerneordforrådet

Blöndals ordbog indeholder som før nævnt 108.000 lemmaer, mens OMI har 56.000 lemmaer. Til trods for ordbøgernes forskellige størrelse foretog vi en sammenligning af opslagsordene med det formål at finde frem til de ord som er permanente i sproget, da en del af ordforrådet er uafhængig af tid (diakroniske ord). Resultatet er noget som man eventuelt kan betegne som kerneordforrådet i islandsk fra ca. 1900 til dagen i dag.

Automatiske kørsler af lemmalisterne viser at 24.000 ord er fælles for begge værker. Dette er et interessant resultat og giver måske et fingerpeg om kernen i det islandske ordforråd. Disse 24.000 ord udgør 43 % af OMI, men derimod kun 22 % af ordforrådet i Blöndals ordbog. OMI indeholder således en betydelig større del af det formodede kerneordforråd end Blöndals ordbog.

I en kort artikel er det ikke muligt at foretage en dybtgående sammenligning af de to ordbøger, og vi begrænser os til et lille udpluk som har det formål at undersøge nogle ændringer i ordforrådet over en periode af hundrede år. Vi har valgt at se nærmere på nogle sammensatte ord i de to ordbøger, for det første ord med endelsen *-fólk* ('-folk, -mennesker') og for det andet ord med førsteleddet *brauð-* ('brød-')

3.2. Sammenligning af *-fólk*

Sidsteleddet *-fólk* indgår aktivt i sammensætninger, både i nutidssproget såvel som i ældre sprog. Her følger de 23 sammensatte ord med endelsen *-fólk*, der er fælles for begge ordbøger:

almúgafólk, alþýðufólk, fyrirfólk, heimafólk, heimilisfólk, huldufólk, kaupafólk, kunningjafólk, kvenfólk, mannfólk, móðurfólk, samferða-

fólk, skyldfólk, sómafólk, starfsfólk, sveitafólk, tengdafólk, utanbæjarfólk, verkafólk, vinafólk, vinnufólk, þjónustufólk, ættfólk

Alle disse ord er en fast bestanddel af det islandske sprog den dag i dag. Det er interessant at Blöndals ordbog derudover har 70 andre sammensætninger med endelsen *-fólk,* mens OMI indeholder yderligere 83 andre sammensætninger (disse ordlister følger dog ikke med artiklen). Det ser således ud til at denne endelse er noget mere produktiv nu end for et hundrede år siden. Det kan eventuelt hænge sammen med øget bevisthed omkring køn at der nu er større tendens til at danne sammensætninger med endelsen *-fólk* '-folk' i stedet for *-menn* '-mænd'. Mange af de fælles ord henviser til sociale relationer eller familiebånd, f.eks. *heimilisfólk* 'personer som tilhører en husstand', *skyldfólk* 'slægtninge', *kunningjafólk* 'bekendte', *vinafólk* 'venner' og *þjónustufólk* 'tjenestefolk'.

Eksempler på ord som ikke er fælles og kun optræder i Blöndals ordbog: *bólufólk* 'mennesker som lider af kopper', *hoffólk* 'hoffolk' og *kirkjufólk* 'kirkegængere'. De tre valgte ord har et arkæisk præg, en moderne islænding ville ikke umiddelbart forstå ordet *hoffólk*.

Ord som kun forekommer i OMI er bl.a. *hálaunafólk* 'højtlønnede personer', *hjúkrunarfólk* 'plejepersonale' og *menntafólk* 'akademikere'. Dette er ganske almindelige ord i moderne islandsk, men for hundrede år siden var de ikke blevet en fast bestanddel af sproget.

I øvrigt forekommer i OMI ord som f.eks. *hjólreiðafólk* 'cyklister' og *handverksfólk* 'kunsthåndværkere', mens en variant af dem står dog i Blöndals ordbog som entalsord: *hjólreiðamaður* 'mandlig cyklist' og *handverksmaður* 'mandlig kunsthåndværker'. Dette genspejler eventuelt kønnenes stilling på den tid.

3.3. Sammenligning af *brauð-*

Brauð- er et meget almindeligt førsteled, som i nogen grad er uafhængigt af tid. Som sidsteled er *-brauð* derimod meget tidsbestemt og beskriver de typer brød som bages og sælges på et givet tidspunkt (f.eks. nutidens *surdejsbrød* og *speltbrød*).

Vi undersøgte førsteleddet *brauð-* i begge ordbøger. Sammenligningen viser at der kun findes 14 fælles sammensætninger i ordbøgerne:

brauðbakstur, brauðbiti, brauðdeig, brauðfæða, brauðfætur, brauðgerð, brauðhleifur, brauðhnífur, brauðkolla, brauðmoli, brauðmylsna, brauðskorpa, brauðsneið, brauðsúpa

Desuden indeholder OMI 12 andre sammensætninger med *brauð-* som første led:

brauðbotn, brauðbretti, brauðform, brauðger, brauðmeti, brauðostur, brauðrasp, brauðrist, brauðstrit, brauðteningur, brauðterta, brauðvél

Der er således ialt 26 sammensatte ord med førsteleddet *brauð-* i OMI. Ordene *brauðrist* 'brødrister' og *brauðvél* 'brødbagemaskine' er selvfølgelig moderne hulsholdningmaskiner som ikke fandtes i nitten hundrede og tyverne.

Blöndals ordbog indeholder 34 andre sammensætninger med *brauð-*:

brauðbakki, brauðbítur, brauðbúð, brauðbökun, brauðdiskur, brauðfæra, brauðgerðarhús, brauðgerðarmaður, brauðgjörð, brauðgrautur, brauðhilla, brauðjurt, brauðkarfa, brauðkassi, brauðkássa, brauðkefli, brauðkringla, brauðlaus, brauðmél, brauðmjöl, brauðmót, brauðofn, brauðreka, brauðsala, brauðsali, brauðselja, brauðskífa, brauðsölubúð, brauðtigl, brauðtunna, brauðvala, brauðvatn, brauðveisla, brauðþurfi

Ialt er der 48 sammensatte ord med førsteleddet *brauð-* i Blöndals ordbog, men kun 14 har fundet vej til OMI. Mange af disse ord er brugbare i moderne islandsk, selv om de ikke blev medtaget i OMI, herunder er mange transparente sammensætninger (*brauðhilla* 'brødhylde', *brauðlaus* 'brødløs, uden brød'). Derimod indeholder listen andre ord som virker fremmede i nutiden, bl.a. *brauðfæra* 'ovnskyder, grissel', *brauðbítur* 'person som lever på andre', *brauðselja* 'kvinde der sælger brød' og *brauðtunna* 'brødtønde'. Dette kommer ikke som en overraskelse da sproget stadig fornys og de fænomener som det beskriver, ændres i tidens løb.

4. Konklusion

Artiklen fremlægger en sammenligning af ordforrådet i to islandske ordbøger, Blöndals ordbog og OMI. Et udvalg af sammensætninger fra de to værker gav os oplysninger om hvilke ord der er fælles for begge ordbøger og hvilke ikke. De ord som ikke er fælles, giver et fingerpeg om sproglig udvikling, undertiden også samfundsmæssig, over et hundrede år, dog

med det forbehold at Blöndals ordbog er betragtelig større end OMI. Ved hjælp af automatiske kørsler af lemmalisterne fandt vi desuden frem til at 24.000 ord er fælles for begge værker, hvilket i sig selv er interessant. Dette kalder på en dybere analyse af de fælles ord: er der hovedsagelig tale om usammensatte ord eller består listen af et stort antal af sammensatte ord (komposita). Disse 24.000 ord kan muligvis betegnes som kerneordforrådet i islandsk, men tilbage står spørgsmålet om hvad der karakteriserer et sprogs kerneordforråd. Formodentlig drejer det sig om ordforråd som har været permanent i sproget over en længere periode.

Litteratur

Bjarnadóttir, Kristín 1997. Allravagn og aðgöngumiðaokrari: um samsett orð í orðabók Blöndals. *Orð og tunga* 3, 61-70.

Blöndal, Sigfús 1920-24. *Íslensk-dönsk orðabók / Islandsk-dansk ordbog.* Reykjavík.

Íslensk-dönsk orðabók (Islandsk-dansk ordbog) digitaliseret <blondal.arnastofnun.is>. Tilgået juni 2022.

Íslensk nútímamálsorðabók. Halldóra Jónsdóttir & Þórdís Úlfarsdóttir (red.). Stofnun Árna Magnússonar í íslenskum fræðum. <islensk-ordabok.is>. Tilgået juni 2022. Forkortet OMI.

Íslenskt textasafn. Úlfarsdóttir, Þórdís (red.). Stofnun Árna Magnússonar í íslenskum fræðum. <corpus.arnastofnun.is/>. Tilgået juni 2022.

Jónsdóttir, Halldóra & Þórdís Úlfarsdóttir 2019. Íslensk nútímamálsorðabók–kjarni tungumálsins. *Orð og tunga* 21, 1-25.

Jónsdóttir, Halldóra & Þórdís Úlfarsdóttir 2020. Omdannelsen af en flersproget til en monolingval ordbog. *Nordiska studier i lexikografi 15. Helsingfors 4-7 juni 2019,* 175-184.

Jónsdóttir Halldóra & Þórdís Úlfarsdóttir 2021. Stafræn gerð Blöndalsorðabókar. *Orð og tunga* 23, 161-166.

Risamálheildin (Gigaword Corpus) 2018. Stofnun Árna Magnússonar í íslenskum fræðum. <malheildir.arnastofnun.is>. Tilgået juni 2022.

Steingrímsson, Steinþór, Sigrún Helgadóttir & Eiríkur Rögnvaldsson 2018. An Icelandic Gigaword Corpus. *Nordiske Studier i Leksikografi 14, Reykjavík 30.5-2.6.2018*, 246-254.

Úlfarsdóttir, Þórdís 2013. ISLEX – norræn margmála orðabók. *Orð og tunga* 15, 41-71.

Sensitive ord i *Det Norske Akademis ordbok*. Utfordringer i diakron leksikografi

Hanne Lauvstad

In 2021–2022 the lexicographers of the *Norwegian Academy Dictionary* (NAOB) have been working on a subproject concerning terms for gender, sexuality, ethnicity (2021), medical conditions and disabilities (2022). The outcome of the project is a revision and update of totally 1371 entries, and the inclusion of totally 49 new entries in the dictionary. This systematic examination and revision has increased the lexicographers' awareness of the subject, and resulted in a more homogenous lexicographical approach. The fact that NAOB is a diachronic dictionary, documenting and describing the vocabulary of a period spanning the last 200 years, and based on literature (fiction and non-fiction), represents special challenges. It has been important not to censor the material, but to formulate the editorial examples, use markers and definitions according to modern norms in a plain, non-activist way, especially to avoid prejudices and stereotypes in all descriptive parts of the dictionary. The lexicographical treatment of older, sensitive terms and their stylistic nuances, as well as the question to what extent the dictionary users are able to fully understand the register labels and the (often utterly concise) definitions has to be considered, which means that a hermeneutic awareness is required when working on this type of subject.

KEYWORDS: sensitive words, diachronical lexicography, practical lexicography, hermeneutics

1. Innledning

Det finnes særskilte språklige domener som har forandret seg ekstra mye, i takt med samfunnsutviklingen for øvrig. Slike domener er f.eks. kjønn, seksualitet, etnisitet, funksjonshemninger og psykiske lidelser. Ordbøkenes beskrivelse av slike domener kan fort virke gammelmodig. Dette er bakgrunnen for et eget prosjekt med to delprosjekter som nylig har vært gjennomført av leksikografene i *Det Norske Akademis ordbok* (NAOB). I 2021 arbeidet redaksjonen med et delprosjekt kalt Sensitive ord i NAOB. Det dreide seg om en leksikografisk undersøkelse og bearbeiding av ord og uttrykksmåter som gjelder kjønn, seksuell legning og etnisk tilhørig-

het. I 2022 er trinn to, med prosjekttittelen Mangfold i NAOB, gjennomført. Det har vært et lignende arbeid med betegnelser for funksjonshemninger og psykiske lidelser.

2. Bakgrunn

Som leksikograf vil man være preget av sin egen tid og sine egne forutsetninger, men man bør unngå eldre tiders stereotypier knyttet til kjønn, etnisitet og seksuell legning i de delene av ordbokartiklene man selv utformer. Dette bør man ha in mente når man arbeider med en dokumentasjonsordbok som dekker språket i de siste 200 årene, slik NAOB gjør. Problemet blir ikke mindre av at NAOB har som utgangspunkt *Norsk Riksmålsordbok* (1937–57, med to tilleggsbind fra 1995). Det finnes dermed et stort nedarvet materiale, både lemmaer, bruksbeskrivelser og redaksjons- og sitateksempler, som bør betraktes med argusøyne. NAOB er basert på litteratur i vid forstand, både skjønnlitteratur og forskjellige typer sakprosa. Disse kildene vises gjennom autentiske eksempler, dvs. sitateksempler, i ordbokartiklene. Ordboken beskriver videre ordforrådet gjennom definisjoner (ordforklaringer) og redaksjonseksempler. Arbeidet med de temaene som er omtalt ovenfor, har gjeldt den leksikografiske beskrivelsen, og generelt den leksikografiske presentasjonen gjennom lemmatisering og lenking.

Leksikografene ved *Svensk ordbok utgiven av Svenska Akademien* (SO) og *Den Danske Ordbog* (DDO) har allerede gjennomført en tilsvarende gjennomgang som den leksikografene i NAOB har arbeidet med. Resultatene er beskrevet i artikler av Petersson og Sköldberg (2020) og Trap-Jensen (2020), begge i LexicoNordica nr. 27. Disse artiklene har vært til nytte og inspirasjon. Riktignok hadde NAOB-redaksjonen som mål å modernisere ordboken på lignende måte som et ledd i det leksikografiske arbeidet som ledet frem til publiseringen av NAOB 21.12.2017, men det var åpenbart for oss i den nåværende NAOB-redaksjonen at det var behov for en mer systematisk gjennomgang av ordbokbasen, og at det ville være nødvendig med tilskudd både av nye lemmaer og nye sitateksempler i ordboken.

Plan Norges ungdomsorganisasjon hadde høsten 2020 aksjonen «Formet av ord» (https://www.plan-norge.no/formet-av-ord), hvor deltagerne tok for seg kjønnsstereotypier i flere norske ordbøker, med særlig vekt på

formuleringer som «en modig mann», «en søt jente» (jf. Nøttveit 2020). Henvendelsen NAOB-redaksjonen fikk fra denne organisasjonen, var også en påminnelse om at det var behov for en oppdatering av NAOB i beskrivelser og omtaler av forhold knyttet til kjønn. Den ønskede endringen ble i dette tilfellet først og fremst gjennomført ved at en del redaksjonseksempler ble revidert med tanke på kjønnsbalansen.

3. Metode

En hovedsak i arbeidet med sensitive ord i NAOB har vært å identifisere de ordbokartiklene som trenger revisjon, og dessuten få sanket nye lemmaer og uttrykksmåter fra de domenene vi har behandlet.

I løpet av prosjektet har NAOB-redaksjonen lagt vekt på å utvide dokumentasjonen av autentisk språkbruk og derfor hentet inn flere sitater fra kvinnelige forfattere og forfattere med flerkulturell bakgrunn. Den viktigste kilden til studiet av ordforrådet generelt og sanking av sitateksempler er NAOBs tekstkorpus og verktøy. Et av disse verktøyene for korpussøk er utviklet i samarbeid med INESS-miljøet ved Universitetet i Bergen og Nasjonalbiblioteket. Nasjonalbibliotekets digitale tjeneste Nettbiblioteket har store mengder litteratur i bokform (skjønnlitteratur og sakprosa), aviser og tidsskrifter tilgjengelig. En del av dette litterære materialet er tatt ut i et eget NAOB-korpus. Dette korpuset er avgrenset til bokmåls- og riksmålstekster, dvs. tekster fra begynnelsen av 1900-tallet til ca. 2000. Vi har fått utviklet en egen søkefunksjon for dette korpuset. I tillegg har vi benyttet et internt NAOB-korpus, som inneholder en del eldre og nyere tekster, som av rettighetshensyn ikke er tilgjengelige for oss via Nettbiblioteket. Disse tekstene har vi fått direkte fra forlagene. Aviser og tidsskrifter er studert og sitert ved hjelp av Nettbibliotekets digitale avis- og tidsskriftsamling og mediearkivet A-tekst (se Nilstun 2023). Dessuten har vi ekserpert (især nyere) litteratur med relevans for de omtalte emneområdene, både skjønnlitteratur og sakprosa, beregnet på så vel voksne som barn og ungdom.

NAOB-redaksjonen har også samarbeidet med leksikografene tilknyttet SO ved Göteborgs universitet i forbindelse med våre delprosjekter Sensitive ord i NAOB og Mangfold i NAOB. Vi har vært så heldige å kunne ta utgangspunkt i lemmalister fra SO med betegnelser for etnisitet, seksuell legning, kjønn og kjønnsidentitet og funksjonshemninger

som har vært aktuelle for NAOB-arbeidet. Vi har valgt norske ekvivalenter til de svenske ordene og assosiert fritt ut fra disse, slik at vi har funnet relevante synonymer, avledninger og sammensetninger. I arbeidet med de sensitive ordene har vi også vurdert innholdet i ordlister laget av aktivistgrupper, som *Den store regnbogeordlista* (2020 [2018]) for LHBT-relaterte ord, utarbeidet av Andrea Rygg Nøttveit i tidsskriftet *Framtida*. I slike tilfeller har en bevissthet om opphavspersonenes aktivistiske perspektiv vært nødvendig. Vi har konsentrert oss om å dokumentere faktisk språkbruk, mens betegnelser som ennå ikke er tatt i bruk, er utelatt. En annen fremgangsmåte for å finne relevant materiale, har vært søk på bestemte fagmarkører i ordbokbasen. I tillegg har vi kunnet finne frem til lemmaer i ordboken med behov for gjennomgang ved å søke trunkert på sisteledd i relevante sammensetninger, som f.eks. *-anstalt*, *-asyl*, *-hjem*. I de omtalte tilfellene dukket det opp betegnelser for behandlingsinstitusjoner som var mer eller mindre foreldede, eller som kunne oppfattes som støtende.

4. Eksempler på behandling av forskjellige domener

La oss se på noen eksempler på ord NAOB-leksikografene har revidert i de to delprosjektene. Ordene gjelder:

- kjønn: Substantivet *hardhaus* var tidligere utelukkende definert som 'hardfør kar', med sitateksempler som dreide seg om menn. Ordet brukes imidlertid like gjerne om kvinner, derfor er NAOB-definisjonen endret til 'hardfør person', og det er lagt til sitateksempler som gjelder kvinner.

- seksualitet: Den eldre definisjonen i NAOB av *svigerdatter* som 'sønns ektefelle' inkluderte ikke personer i likekjønnede ekteskap. Her har både språklige og utenomspråklige forhold (ekteskapsloven og øvrige samfunnsforhold) endret seg, og det har ført til at definisjonen har måttet endres i NAOB til 'kone til (eller samboer med) ens sønn eller datter' (ordet *svigersønn* er tilsvarende revidert).

- etnisitet: Et ord som *eskimo* har fått revidert ordhistorisk redegjørelse, bruks- og stilmarkører og nye sitater. Ordet *neger* var allerede grundig gjennomgått, men fikk reviderte bruks- og stilmarkører.

- psykiske lidelser: Mange av de relevante fagtermene ble gjennomgått før publiseringen av NAOB i 2017, så revisjonen som trengtes, var mindre enn ved mange andre domener. Noen endringer er allikevel gjort, f.eks. har den psykiatriske fagtermen *angsthysteri* (tidligere definert som 'sterk angstnevrose') fått revidert definisjonen til 'fellesbetegnelse for fobier' og er blitt markert som «foreldet», mens en annen term fra samme fagområde, *borderline*, har fått reviderte ordhistoriske opplysninger og blitt utstyrt med et sitateksempel.

- funksjonshemninger: En rekke definisjoner og redaksjonseksempler er gjennomgått og endret. Disse gjelder både betegnelser for funksjonshemninger og andre betegnelser som tilhører domenet, f.eks. betegnelser for hjelpemidler som *rullator* (hvor definisjonen er gjennomgått og sitateksempler lagt til) og *punktskrift* (hvor et sitateksempel er tilføyd).

- artikler som gjelder betegnelser med relevans for flere domener, som betegnelser for bestemte typer institusjoner, er også revidert, f.eks. er en del sammensetninger på *-anstalt* og *-asyl* markert som foreldede.

5. Problemstillinger

- Ved behandlingen av de sensitive ordene er det viktig å vurdere både *avsenderintensjonen* og *mottagerforståelsen* av de betegnelsene som blir vurdert og beskrevet på nytt. Dette aspektet er grundig omtalt av f.eks. Petersson og Sköldberg (2020) og Trap-Jensen (2020).

- Det spesielle i vår sammenheng har vært det man kunne kalle *det diakrone problem*. Siden NAOB er en dokumentasjonsordbok som dekker de siste par hundre årene, må vi vise utviklingen i bruk av

ord som nå kan være sensitive i en eller annen forstand, men som tidligere ikke nødvendigvis ble oppfattet eller anvendt slik. Vi får altså et system som kan beskrives slik:

historisk avsender	(samtidig) historisk mottager
historisk avsender	moderne mottager
moderne avsender	moderne mottager

- Et annet viktig moment i arbeidet med de sensitive ordene er grenseoppgangen mellom fagspråk og allmennspråk. Man må også tydeliggjøre hvilke fagtermer som er foreldede, noe som iblant kan være krevende. I noen tilfeller er tidligere fagtermer fortsatt i bruk i mer uformell språkbruk, noe det har vært viktig å fremheve i ordbokartiklene.

Hans-Georg Gadamer (1972 [1960]) har pekt på de hermeneutiske problemer som følger med lesningen av historiske tekster og lansert begrepet *forståelseshorisont*. Gadamer legger vekt på at både den som er avsender og den som er mottager av et budskap, er preget av sine sosiale og/eller historiske forutsetninger, dvs. at de to partene kan ha høyst forskjellig forståelseshorisont. Også mennesker som lever samtidig, kan ha ulike forutsetninger for å forstå en tekst. Det at en person (en mottager) tilegner seg en tekst, innebærer en fortolkning. Denne fortolkningen er ifølge Gadamer ikke bare subjektiv, men preget av den sosiale og historiske konteksten leseren befinner seg i. Med Gadamers ord er leseren preget av sin «Vorurteil» ('fordom', her: 'forforståelse') (Gadamer 1972 [1960]:282). Situasjonen kan beskrives slik: En historisk avsender har skrevet en tekst som inneholder et budskap. Dette budskapet kan oppfattes av en samtidig, historisk mottager, eller av en moderne mottager, noe som innebærer diakront og/eller synkront forskjellige typer forforståelse. En moderne avsender formidler et budskap til en samtidig mottager, også her kan fortolkningen variere fordi mottageren kan ha en annen forforståelse enn avsenderen av teksten. Disse faktorene er det praktisk å være bevisst når

man, som i NAOB, har litterære kilder til ordboken som stammer fra forskjellige perioder, og som via ordboken skal presenteres for moderne ordbokbrukere (med sin type forforståelse) på en leksikografisk forsvarlig måte. Det er altså viktig å ha en hermeneutisk bevissthet som leksikograf.

Som eksempler vil jeg ta ordene *raseansikt* og *renraset*.

raseansikt substantiv

BØYNING et

BETYDNING OG BRUK

NÅ SJELDEN, OPPFATTES PROVOSERENDE ansikt med rase | jf. rasehode

SITATER

- *hun kunde staa ansigt til ansigt med de mange, høi og vakker med det sterke raceansigt* (Kristian Elster d.y. *Bonde Veirskjæg* 66 1930)
- *det første trekk som slo én i hans skarpt skårne raseansikt [var] et uttrykk av tretthet og vemod* (Lorentz Eckhoff *Førerne i vår tids franske litteratur* 36 1928)

FIGUR 1. Artikkelen *raseansikt* i NAOB.

I sitatene i artikkelen *raseansikt* ser det ut til at avsenderne har brukt ordet i positiv betydning. I dag, etter at verden har erfart nazistenes ideologi og dens følger, ville det være umulig å bruke ordet på samme måte. Derfor er det lagt til bruks- og stilmarkører.

renraset adjektiv

BØYNING renrasede

ETYMOLOGI annet ledd avledet av rase

BETYDNING OG BRUK

1 LANDBRUK, VETERINÆRFAG av ren, ublandet rase | jf. avl

SITAT

- *[valpen] er ikke renraset, hadde de sagt, men i moren er det jo mye retriever, det vet vi i hvert fall* (Torun Lian *Undrene i vår familie* LBK 2008)

2 OM PERSON, KAN OPPFATTES STØTENDE ELLER PROVOSERENDE som har forfedre med opprinnelse i én etnisk gruppe, befolkningsgruppe

SITATER

- IRONISK, POLEMISK *tante Eileen var meget velhavende. Og hun var renraset jødinne. [Den nazistiske] Erling [Bjørnson] satte jord og helvete i sving for å få kloen i pengene* (B.A. Bjørnson-Langen *Aulestad tur-retur* 126 1981)
- IRONISK, POLEMISK *vi lærte visst det også, at nordmennene var den reneste nordiske rasen. Atter en evig løgn ... nei, vi er nok dessverre intet renraset land* (Sigurd Hoel *Tanker fra mange tider* 218 1948) | fra essayet «Bør det bo folk i Norge?» (1946)
- *Cholon sa at en araber var en araber, renraset, i motsetning til dette rasket som kalte seg parisere og hadde silkeormer mellom beina og en syfilitisk sekt i hodet* (Ola Bauer *Magenta* LBK 1997)

FIGUR 2. Artikkelen *renraset* i NAOB.

Ordet *renraset* i betydning 1, om dyreavl, må sies å være uproblematisk i vår sammenheng. Betydning 2 er mer problematisk, pga. eldre tiders raseteori og rasisme. De første to sitatene var i utgangspunktet nokså knappe, og det var umulig å lese avsenderintensjonen ut av dem slik de sto. En videre kontekst måtte trekkes inn i vurderingen. Det var altså nødvendig å gå direkte til kildene og gjengi mer tekst i sitatene og dessuten legge til utdypende forklaringer og nødvendige stilmarkører. Sitatet av Bjørnstjerne Albert Bjørnson-Langen, Bjørnstjerne Bjørnsons barnebarn, er en del av en polemisk omtale av onkelen Erling Bjørnson, som var medlem av det norske nazipartiet Nasjonal Samling. Teksten inneholder fortelleteknikken fri indirekte diskurs (med referanse til tredjeperson og verb i fortidsform), som ironisk gjengir nazistens synsvinkel. Sitatet er slik sett problematisk når det er tatt ut av sin opprinnelige kontekst, men samtidig er det et interessant vitnesbyrd om etterkrigstidens oppgjør med ordet *renraset* og de holdninger som var blitt lagt i det. Sitatet av Sigurd Hoel inneholder likeledes en ironisk polemisering mot nazistisk raseoppfatning, noe som på samme måte er markert i ordboken.

Grenseoppgangen mellom fagspråk og allmennspråk er som nevnt en del av problemstillingen. Betegnelser kan ha eksistert i uformell språkbruk og enten være tidligere akseptable fagtermer, som *sinnssyk*, eller de kan aldri ha vært akseptable i faglig eller mer seriøs sammenheng. Dette gjelder et ord som *gal*, som typisk tilhører uformelt språk. Det kan iblant være vanskelig å vurdere fortidens betegnelser, som f.eks. adjektivet *forrykt* (og det tilhørende substantivet *forrykthet*). Man må avklare hvorvidt disse betegnelsene tidligere kan ha hatt faglige konnotasjoner og slik sett ha vært mindre ladede enn i dag. Det ser ut til at ordene primært har hatt en flytende dagligspråklig referanse i retning av henholdsvis 'gal (også sinnssyk)' og 'galskap', men at de også tidligere har vært brukt i en avgrenset betydning: adjektivet *forrykt* i betydningen 'som lider av systematiske vrangforestillinger, især paranoia' og substantivet *forrykthet* i betydningen 'det å lide av systematiske vrangforestillinger (paranoid psykose)'. *Store norsk leksikon* (SNL) er noe vagt når det gjelder den moderne bruken og betydningen av disse ordene, mens *Svenska Akademiens ordbok* (SAOB) har (riktignok i annen rekkefølge og med flere detaljer) en tydeliggjøring à la den vi har bestemt oss for i NAOB, en inndeling med tydelig skille mellom allmennspråklig og fagspråklig betydning: 1 (i dagligtale) 'gal; sinnssyk; vanvittig'; 2 (psykologi, psykiatri, foreldet som fagterm)

'som lider av systematiske vrangforestillinger, især paranoia'. Substantivet *forrykthet* er behandlet på lignende måte.

Vurderingen av eldre avsenderintensjon og mottageroppfatning må altså først og fremst gjøres ved hjelp av den kontekst et ord forekommer i, og da er det en fordel å ha flere belegg å bygge på. I slike tilfeller kan man møysommelig skaffe seg et inntrykk av hva som har vært et ords alminnelige konnotasjoner i en bestemt periode. Men dersom betegnelsen er lavfrekvent, blir dette vanskelig. En viktig kilde er andre historiske ordbøker; i NAOB har vi f.eks. god nytte av ODS og SAOB, når det gjelder dialektale ord *Norsk Ordbok*, iblant også andre historiske ordbøker. Selv om ordbøkene dekker forskjellige språkområder, finnes det som regel fellestrekk i ordhistorien, ikke minst mellom de skandinaviske språkene. Fagtermer kontrolleres også mot leksika og andre terminologiske kilder, som fagordbøker og termlister. Dersom man mistenker at termene er foreldede, kan eldre oppslagsverk eller faglitteratur bidra til å belyse forholdene.

6. Endringer av ordbokens mikro- og makrostruktur

Arbeidet med sensitive ord i NAOB har fått konsekvenser for både makro- og mikrostrukturen i ordboken.

- Nye ord er blitt tatt inn i NAOB, f.eks. *antisiganisme*, *cis* (adjektiv), *cismann*, *ciskvinne*, *funkofobi*, *medforelder*, *transfobi* og *transfobisk*.

- Definisjoner er blitt endret: den endrede definisjonen av *svigerdatter* er allerede omtalt. Et annet eksempel, som gjelder fremstillingen av kjønn i ordboken, er *sex appeal*. Tidligere var det definert som 'evne (særl. hos kvinne) til å virke erotisk tiltrekkende'. Dette er endret til 'evne til å virke seksuelt tiltrekkende; erotisk utstråling', og det er tatt inn sitateksempler som også gjelder menn.

- Betydningsinndelingen er blitt revidert flere steder, og underbetydninger i noen tilfeller fremhevet. I tillegg er mer subtile språktrekk blitt avdekket under arbeidet og har fått en tydeligere presentasjon. Som eksempel kan nevnes at adjektiver som *blind*, *døv*, *funksjonshemmet* er alminnelige i substantivert form som gruppebetegnel-

ser («de blinde», «de døve», «funksjonshemmede». Dette er blitt fremhevet i NAOB ved at de substantiverte adjektivene har fått egne underbetydninger med definisjoner og sitateksempler. Denne leksikografiske tydeliggjøringen kan synliggjøre bestemte grupper i samfunnet.

- Bruksmarkører er endret en rekke steder i ordboken. Forskjellen mellom avsenderintensjon, i stilmarkører som «nedsettende», «polemisk», og mottagerperspektiv, i formuleringer som «kan oppfattes støtende», er gjennomgått og revidert. Oppslagsordet *indianer* har f.eks. nå fått bruksmarkøren «kan oppfattes støtende». Tidligere stod ordet uten stilmarkør.

- Redaksjonseksemplene er revidert flere steder. NAOBs redaksjonseksempler var tidligere mannsdominerte. I en rekke tilfeller er pronomenet i eksempelet endret fra «han» til «hun» (og ved språkeksempler som viser til flere personer, til flertallspronomener). Særlig viktig har det vært å ha med redaksjonseksempler som bryter med stereotype oppfatninger av ord som *sterk* («Pippi Langstrømpe – verdens sterkeste jente») eller *modig* («en modig mann» er byttet ut med «hun/han er modig»).

- Sitateksempler som viser flere aspekter ved oppslagsordet eller en ny type kontekst for oppslagsordet, er tatt inn, f.eks. under *tøff*: «hun var en tøff dame. I unge år hadde hun trent kampsport» (Tom Egeland: *Kongen* (2020, side 75)), og under oppslagsordet *døv*: «på Stortinget var den døve kopisten [Lars] Havstad kjent som et arbeidsjern» (Hilde Diesen: *Talegaven* (2021, side 100)).

- Lenkingen mellom ordbokartiklene er forandret, slik at lenkene nå konsekvent fører fra mindre gangbare betegnelser til det som anses som mer moderne og umarkerte betegnelser, f.eks. er det nå lenket fra *eskimo* til *inuitt*, men ikke omvendt, slik det opprinnelig var gjort. Et annet eksempel er at *autisme* nå er lenket til *autismespekterforstyrrelse*, som forholdsvis nylig ser ut til å være blitt den gangbare fagtermen.

Hovedhensikten med delprosjektet var en revisjon av eksisterende ordbokartikler i NAOB. Men prosjektarbeidet medførte også at noen nye ordbokartikler ble utarbeidet.

De tallmessige resultatene etter 2021-delprosjektet er at 834 ordbokartikler er gjennomgått og revidert ved behov, mens 26 nye ordbokartikler er utarbeidet. I delprosjektet i 2022 er 537 ordbokartikler gjennomgått og revidert ved behov, mens 23 nye ordbokartikler er utarbeidet.

7. Konklusjon

Gjennomgangen av NAOBs artikler med tanke på spesielle semantiske grupper av sensitive ord eller bestemte domener som særlig trenger leksikografisk gjennomgang og modernisering, har vist seg å være nyttig. En sammenhengende gjennomgang av ordbokens språklige presentasjoner av domenene kjønn, seksualitet, etnisitet, funksjonshemninger og psykiske lidelser har økt den leksikografiske bevisstheten og modernisert ordboken på disse punktene. Gjennomgangen har gjort det lettere å identifisere svakheter i den eksisterende fremstillingen i ordboken. Det har vært viktig å ivareta NAOBs dokumenterende og historiske materiale, men samtidig sikre at presentasjonen i form av definisjoner, bruksmarkører, redaksjonseksempler og lenking gjenspeiler moderne normer. En systematisk gjennomgang av bestemte emner fører til at man også får øye på de mer subtile, indirekte signalene i ordbokbasen og kan gjøre nødvendige endringer. Med disse grepene kan man gjøre ordboken mer moderne og relevant for dagens ordbokbrukere.

Litteratur

Ordbøker, ordlister og leksika

DO = *Den Danske Ordbog*. København: Det Danske Sprog- og Litteraturselskab. I: <ordnet.dk/ddo>. Nedlastet februar 2023.

Den store regnbogeordlista (2020 [2018]), utarbeidet av Andrea Rygg Nøttveit, *Framtida*: < framtida.no/2018/09/14/regnbogeordlista-anno-2018>. [Oppdatert 03.02.2020.] Nedlastet februar 2023.

NAOB = *Det Norske Akademis ordbok*. Oslo: Det Norske Akademi for Språk og Litteratur. I: <naob.no>. Nedlastet februar 2023.

Norsk Ordbok. Bergen: Universitetet i Bergen/Samlaget. <alfa.norsk-ordbok.no/>. Nedlastet februar 2023.

ODS = *Ordbog over det danske Sprog*. København: Det Danske Sprog- og Litteraturselskab. I: <ordnet.dk/ods>. Nedlastet februar 2023.

SAOB = *Svenska Akademiens ordbok*. Lund: Svenska Akademien. I: <saob.se/>. Nedlastet februar 2023.

SO = *Svensk ordbok utgiven av Svenska Akademien*. Göteborg: Svenska Akademien. I: <svenska.se>. Nedlastet februar 2023.

SNL = *Store norske leksikon*. I: <snl.no>. Oslo: Store norske leksikon. Nedlastet februar 2023.

Annen litteratur

Gadamer, Hans-Georg 1972 [1960]: *Wahrheit und Methode. Grudzüge einer philosophischen Hermeneutik*. 3. utvidede opplag. Tübingen.

Nilstun, Carina 2023. Hvordan holde tritt med tiden i en historisk ordbok som også beskriver samtiden? I: Holmer, Louise et al. (red.), *Nordiska studier i lexikografi 16*. Lund & Göteborg: Nordiska föreningen för lexikografi, 257–265.

Nøttveit, Andrea Rygg 2020: Kjønn i ordbøkene, *Framtida* 28.12.2020. Oslo. <https://framtida.no/2020/10/14/kjonn-i-ordbokene>. Nedlastet september 2022.

Petersson, Stellan & Emma Sköldberg 2020: Beskriva utan att diskriminera. Representation av könsidentitet och sexuell läggning i *Svensk ordbok*. *LexicoNordica* 27, 97–115.

Trap-Jensen, Lars 2020: Inklusion eller mindretalsdiktatur? Om politisk korrekthed, minoritetshensyn og leksikografisk deskriptivisme: *LexicoNordica* 27, 137–156.

Teckenspråkslexikografi – utmaningar i en annan modalitet

Johanna Mesch, Elisabet Eir Cortes, Thomas Björkstrand, Nikolaus R. Kankkonen, Joel Bäckström & Patrick Hansson

Swedish Sign Language Dictionary, first published in 2008, was updated to a new version in May 2022. This version has improved search functions and user interface. The lexical database has over 24,000 unique signs and 6,600 example sentences and is being continuously updated with new signs and examples. Each entry contains rich information in the form of video and photos demonstrating the sign, as well as in text and sign transcription describing the execution of the sign. For many signs there is additional information, for instance on the use of the sign in context or on the sign's origin. The different search paths that have been developed are based on the structure of the signs and on different subject areas. Swedish Sign Language Dictionary is connected to the Swedish Sign Language Corpus, which currently contains approximately 190,000 occurrences of signs. This collaboration takes place via the STS-korpus, a web-based tool for the presentation of signs in natural language use. The collaboration between these two language resources is also concerned with lexicographic issues.

NYCKELORD: teckenspråkslexikon, bimodalt-tvåspråkigt lexikon svenskt teckenspråk-svenska, rörliga bilder, samverkan mellan lexikon och korpus, svenskt teckenspråk

1. Introduktion

En viktig aspekt av varje språkresurs är att den är representativ för det eller de språk som den täcker. På grund av detta fungerar resurserna även som språkdokumentation. Teckenspråkslexikon är nödvändiga språkresurser för att tillgodose behoven hos teckenspråkstolkar, teckenspråkslärare, studenter, döva, personer med särskilda behov, forskare och andra användare av teckenspråk, men inte minst för att höja teckenspråkets ställning (McKee & Vale 2017). I denna artikel presenterar vi följande: a) konstruktion av teckenspråkslexikon med teckenposter innehållande teckendemonstration och tillhörande information, utveckling av olika sökfunktioner, insamling av tecken, samverkan mellan lexikon och kor-

pus samt publicering i olika applikationer; och b) lexikografiska frågeställningar som lemma, fonologiska varianter och sammansättningar.

2. Historik över lexikon för svenskt teckenspråk

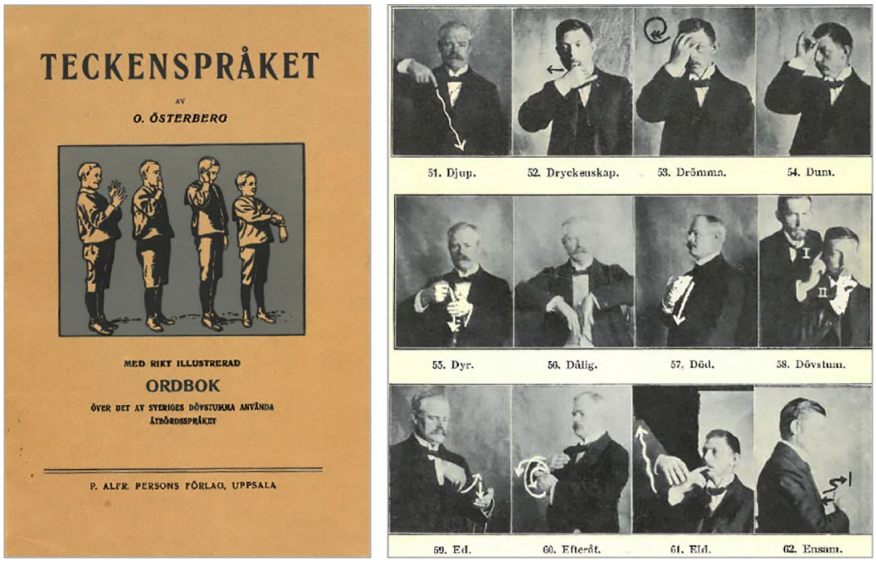
Det första svenska teckenspråkslexikonet i bokformat gavs ut 1916 i skuggan av oralismtiden[1] (Österberg 1916), se figur 1. Det blev en torka fram tills 1960-talet, då flera publikationer kom ut under åren 1960–1978. Den senaste ordboken i tryckt bokformat gavs ut av Svenska Dövas Riksförbund 1997 efter tio års arbete (Svenskt teckenspråkslexikon 1997). I bokform gavs teckenbeskrivningar i form av skriven svenska och fotografier, vilket inte är tillräckligt för att kunna förstå hur tecken utförs. Rörliga bilder är den bästa visningsformen. Det lexikografiska arbetet med svenskt teckenspråk (STS) initierades 1988 av Avdelningen för teckenspråk, Institutionen för lingvistik vid Stockholms universitet. År 2001 gick det första lexikonet från projektet upp på nätet, *Digital version av Svenskt teckenspråkslexikon*, som innehöll 3 132 teckenuppslag. Denna ordbok innehåller rörliga bilder för tecken hämtade från det tryckta lexikonet från Sveriges Dövas Riksförbund. Det stod snart klart att lexikografiska kriterier i den tryckta ordboken gjorde den till en begränsad resurs, bland annat kopplingen till homonymer och ämnesområden. Behov fanns för en mer flexibel ”ordbok” för olika ändamål. I Tabell 1 presenteras produktion av teckenspråkslexikon genom åren.

1 Oralism, eller talmetoden, är en ideologi och praktik som förespråkar kommunikation baserad enbart på tal. Den dominerade dövundervisningen i Sverige (och övriga västvärlden) fram till 70-talet. Talmetoden innebar att döva barn inte fick använda teckenspråk utan skulle lära sig att ljuda orden och läsa på läppar.

TABELL 1. Teckenspråkslexikon som har publicerats genom åren.

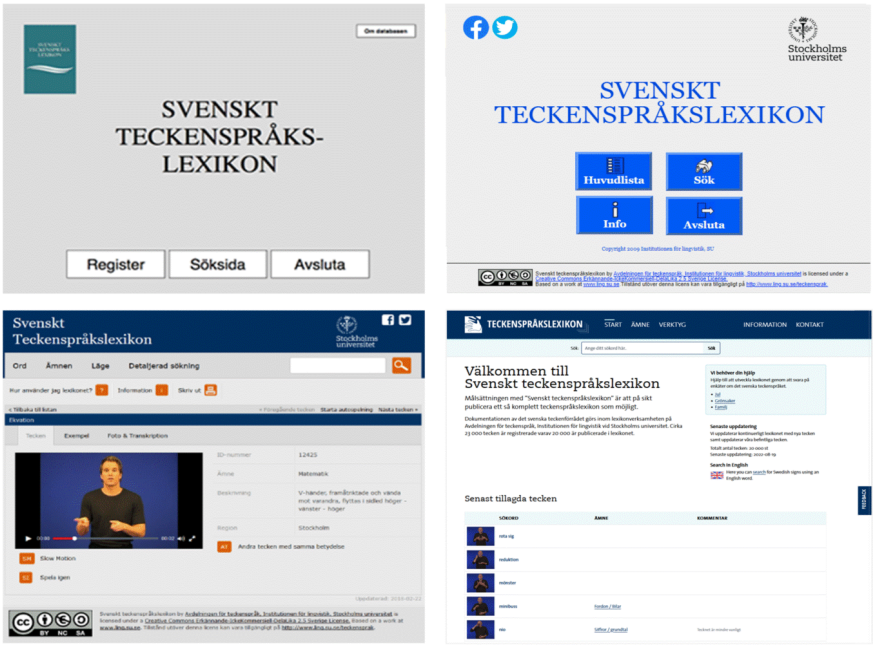
Titel	Publikationsår	Författare/producent	Lexikontyp/form
Teckenspråket	1916	Österberg, Oskar	Bokform
Ordbok över de dövas åtbördsspråk	1960	Kyrkliga nämnden för döva, Stockholm	Bokform
Teckenspråk för döva	1968	Skolöverstyrelsen	Bokform
Teckenordbok	1972	Sveriges dövas riksförbund	Bokform
Teckenboken	1978	Sveriges dövas riksförbund	Bokform
Svenskt teckenspråkslexikon	1997	Sveriges dövas riksförbund	Bokform
Teckenspråk till vardags	2000	Sveriges dövas riksförbund	Bokform
Tecken inom området idrott	2003	Hedberg & Mesch/ Sveriges dövas riksförbund	Bokform, ämnesspecifika lexikon
Digitalt teckenspråkslexikon	2001	Stockholms universitet	Digital form, CD; nio uppdelade skivor – version 1.0
Tecken för Matematiska begrepp	2004	Stockholms universitet	Digital form, CD, ämnesspecifikt lexikon
Tecken för Juridiska begrepp	2004	Stockholms universitet	Digital form, CD, ämnesspecifikt lexikon
Tecken inom Bridge	2004	Stockholms universitet	Digital form, CD, ämnesspecifikt lexikon
Tecken för begrepp inom språkvetenskap	2005	Stockholms universitet	Digital form, CD, ämnesspecifikt lexikon
Kyrkliga tecken	2006	Stockholms universitet	Digital form, CD, ämnesspecifikt lexikon
Tecken för svenska landskap och orter	2006	Stockholms universitet	Digital form, CD, ämnesspecifikt lexikon
Svenskt teckenspråkslexikon	2004	Stockholms universitet	Digital form, DVD – version 1.1
Svenskt teckenspråkslexikon	2008	Stockholms universitet	Webbversion (version 1)
Svenskt teckenspråkslexikon	2014	Stockholms universitet	Webbversion (version 2)
Svenskt teckenspråkslexikon	2022	Stockholms universitet	Webbversion (version 3)

Figur 1 ger en glimt av hur det allra första lexikonet för svenskt teckenspråk, *Teckenspråket, med rikt illustrerad ordbok över det av Sveriges dövstumma använda åtbördsspråket* (Österberg 1916), ser ut. Tecken från detta första lexikon finns som 415 teckenposter i Svenskt teckenspråkslexikon (2022), där de illustreras med de ursprungliga svartvita bilderna/ritningarna från Österbergs ordbok, men också med moderna videoinspelningar som har sepiafärgad bakgrund för att markera att det rör sig om tecken från detta första lexikon.



FIGUR 1. Det första teckenspråkslexikonet av Österberg (1916). Bokens framsida och exempel på teckenillustrationer.

I figuren som följer, figur 2, syns skärmbilder från de olika digitala versionerna av Svenskt teckenspråkslexikon. Bilden nederst t.v. visar innehållet i en teckenpost, (se *Konstruktion och lexikografiskt arbete* här nedan).

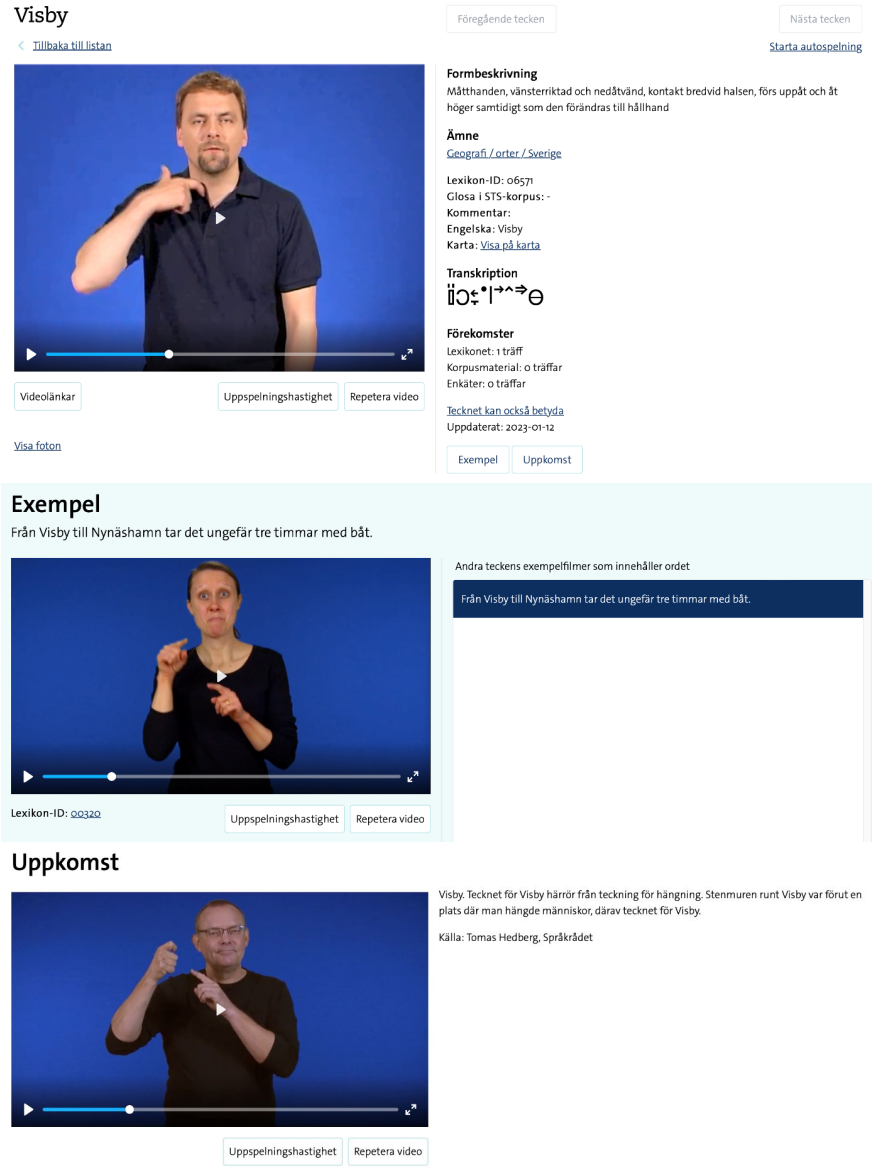


FIGUR 2. Förstasidorna i digitala versioner av Svenskt teckenspråkslexikon. Övre rad från vänster: digital form 2004, webbversion 2008. Undre rad från vänster: webbversion 2014 (visar innehållet i en teckenpost), webbversion 2022.

3. Konstruktion och lexikografiskt arbete

Databasen för Svenskt teckenspråkslexikon består av över 24 000 tecken varav 20 000 är publicerade (augusti 2022). Översättning av tecken till svenska tjänar som grund för s.k. teckenposter (se nästa paragraf), vilket dock inte innebär att varje tecken har en motsvarighet i svenskan. Översättningen till svensk text gör lexikonet bimodalt-tvåspråkigt. Databasen skapades i form av databasfiler i FileMaker Pro, som är ett databasverktyg med inbyggd webbpublicering. Den inbyggda webbpubliceringen används dock inte längre, utan hemsidan körs mot databasverktygets XML-verktyg och styrs direkt från en av lexikonets servrar. Framtagningsarbetet av databasen, som skapades redan i slutet av 90-talet, pågår ännu idag och dess gränssnitt har varit under utveckling sedan dess. Utvecklingsarbetet syftar till att göra det webbaserade lexikonet mer användarvänligt och användbart på olika plattformar. Lexikonteamets gedigna erfarenhet och språkkunskaper är oumbärliga i arbetet med utveckling, uppdatering och

förbättring av lexikonet. Hemsidan för Svenskt teckenspråkslexikon är <https://teckenspraklexikon.su.se/>.



FIGUR 3. Teckenpost för tecknet VISBY. Överst: video med tecknet. T.h. om videon finns formbeskrivning, ämne, transkription, lexikon-ID och annan information. I mitten: video som visar hur tecknet kan användas i en mening. Underst: video med förklaring till tecknets uppkomst.

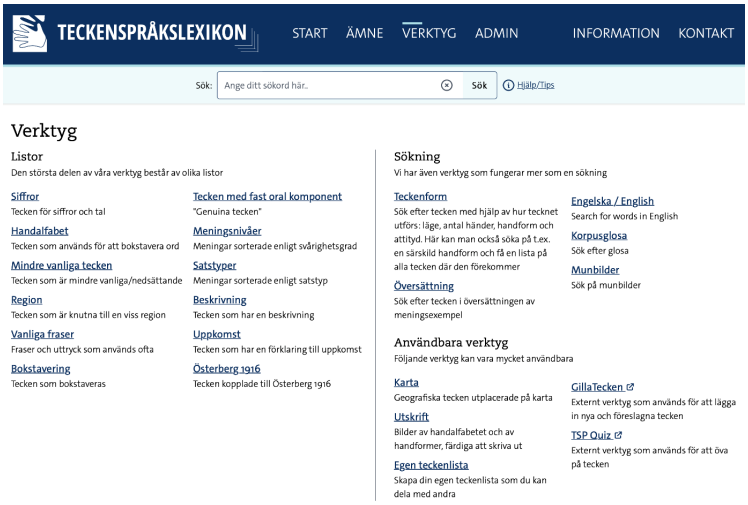
Varje teckenpost har ett unikt ID-nummer. Alla teckenposter innehåller teckendemonstration, svensk översättning, formaliserad formbeskrivning (handform, orientering, läge, rörelse), svenska översättningar och teckentranskription baserad på Bergman och Björkstrands (2015) transkriptionssymboler, och där det är relevant, interna länkar till fonologiskt eller semantiskt ekvivalenta tecken – d.v.s. homonymer och synonymer. För vissa tecken finns kommentarer med information om t.ex. hur vanligt tecknet är, om tecknet kan uppfattas som kränkande, hur tecknet används, eller om tecknets uppkomst (etymologi). Många teckenposter innehåller dessutom användningsexempel, d.v.s. meningar som visar tecknet i kontext.

Figur 3 visar en teckenpost i lexikonet för tecknet VISBY. Teckenposten innehåller, förutom beskrivning av tecknet, användningsexempel och berättelse om tecknets uppkomst.

Lexikografiska frågor är väsentliga i utvecklingsarbetet. Exempel på dessa frågor är vilka tecken som är ”tillåtna” i lexikonet, hur definiera vad som är ett lexikalt tecken, s.k. kärntecken, och vad som är lemma, d.v.s. grundformen av ett ord/uttryck. Val utifrån manuella komponenter möjliggör sökning baserad på teckenform, medan val av teckenbegrepp utifrån begreppsmässiga, semantiska grunder inte är kopplade till det manuella lemmat (Fenlon et al. 2015; McKee & Vale 2017; Schermer 2006). Ytterligare exempel på det utvecklingsarbete som nu pågår är införlivandet av sammansättningar, lexikaliserade avbildande tecken och bokstaveringar i lexikonets befintliga struktur. Jämfört med det lexikografiska arbetet för talade språk, är motsvarande arbete för teckenspråkslexikon mer komplicerat på grund av videomaterial och sökfunktioner.

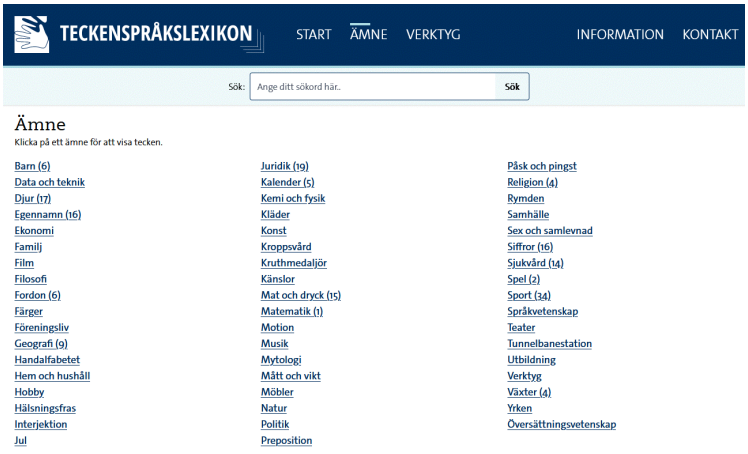
4. Sökvägar

I Svenskt teckenspråkslexikon har verktyg skapats för att kunna söka efter tecken. Verktygen erbjuder flera olika sökvägar: sökning med svenska ord, sökning efter tecken i den svenska översättningen av exempelmeningar, andra betydelser, alternativa tecken, handform, ämnesområde, översättning till engelska, m.m., se figur 4. Här finns även möjlighet att söka efter tecknet i teckenspråkskorpusen.



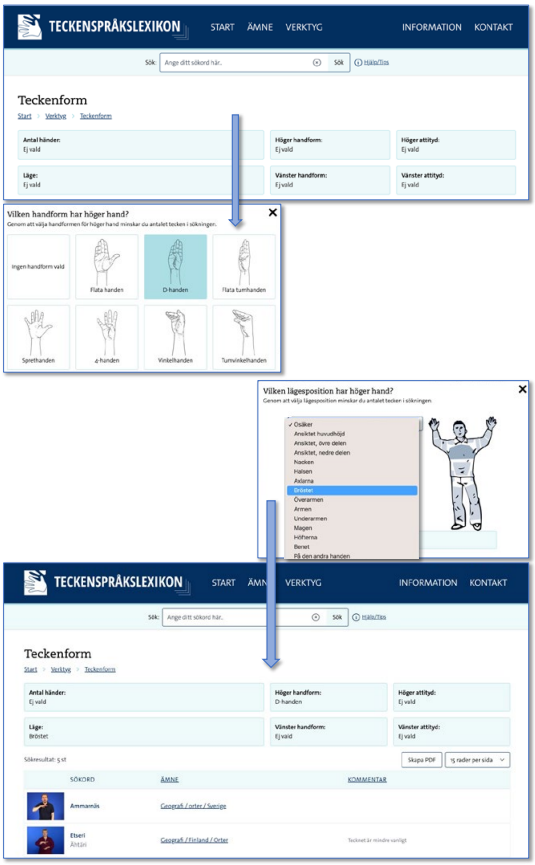
FIGUR 4. Lexikonets verktyg. Här finns olika sökvägar tillgängliga, baserade på t.ex. tecknets olika egenskaper. Sökruta för sökning med svenska ord syns i figurens övre del.

Det vanligaste sökförfarandet är att skriva ett svenskt ord i en sökruta, se figurer 4 och 5. Detta innebär dock inte att varje tecken nödvändigtvis motsvarar ett ord i svenska, utan här söks tecken med en ekvivalent betydelse på svenska. Ett annat vanligt sätt att söka tecken är att gå in i ämneskategorier och välja något ämne, till exempel sjukvård, eller underkategorier för respektive ämne, se figur 5.



FIGUR 5. Sökning via ämneskategorier.
Siffror inom parentes indikerar antal underkategorier.

Unikt för teckenspråkslexikonet är att kunna söka på teckenform. Figur 6 visar hur en sådan sökning kan gå till. Verktyget ger möjlighet att specificera antal händer, handform, attityd (handens/händernas riktning och vridning), samt läge (var på/vid kroppen handen befinner sig). Sökning på teckenform kan göras även om användaren inte vet, eller är osäker på, någon av sökparametrarna.



FIGUR 6. Exempel på sökfunktionen teckenform.
Överst: startläge utan specifikationer. Mitten: val av handens form och läge.
Underst: sökresultatet, en lista med tecken där vald handform och läge används.

5. Insamling av tecken och ”crowdsourcing”

Det självklara valet av källa för insamling av tecken för lexikonet var från början *Svenskt teckenspråkslexikon* (bokformat; Sveriges Dövas Riks-

förbund 1997), som vid den tiden var det största sammanhållna teckenspråksmaterialet som fanns, redan insamlat, beskrivet och katalogiserat. Genom åren har flera insamlingskällor tillkommit. Gemensamt för dessa är att de bygger på ”crowdsourcing”, eller gräsrotsbaserad problemlösning, d.v.s. utveckling och förbättring av lexikonet sker i samarbete med döv- och teckenspråkssamhället. Bidrag från den teckenspråkiga allmänheten är en naturlig och rik källa till teckenspråk från språkanvändare i hela Sverige, varav många är språkvetenskapligt naiva, d.v.s. inte påverkade av ett akademiskt språk eller analytiskt förhållande till språk.

Den mest produktiva av dessa samarbetskällor är de s.k. exempelmeningarna (se figur 3, mellersta bilden). Dessa meningar skapas av aktörer som har STS som första språk. De får se en stillbild av utförandet av ett visst tecken, vilket skapar associationer till möjliga tecken och betydelser. Därefter får aktörerna ta fram sina egna användningsexempel i form av meningar som innehåller tecknet i fråga. Förutom användar- och kontextberoende variation av ett och samma tecken, kan dessa exempelmeningar ge upphov till nya teckenposter i lexikonet. Lexikonteamet arbetar kontinuerligt med att ta fram nya tecken från de 6 600 exempelmeningar som för närvarande finns i lexikonet.

Lexikonets enkäter är ytterligare en viktig källa för nya tecken. Tecken, som redan finns i lexikonet, visas för teckenspråksanvändare som kan antingen rösta på tecknet (bekräfta att de använder det) eller ange ett alternativt tecken. Förutom potentiella nya tecken ger denna källa information om hur vanligt förekommande ett tecken är.

Facebookgruppen Teckenspråkslexikon <www.facebook.com/groups/848909041806827> startades i oktober 2014 och administreras av lexikonteamet. Gruppen innehåller ca 6 100 medlemmar (i september 2022), som interagerar på olika sätt och diskuterar teckenanvändning, lexikal variation och teckenbildning. Under åren 2016 till 2018 skrevs det 593 inlägg och 5 817 interaktioner ägde rum. Detta inkluderar frågor om tecken, till exempel ”Hur tecknar du X?” (53 %), ”Finns det ett tecken för person Y?” (20 %), allmänna frågor om användning (10 %) och om etymologi (<0,2 %) av specifika tecken (Riemer Kankkonen et al. 2018). Tecken som föreslås i Facebookgruppen tas upp i lexikonteamets diskussioner.

En fjärde källa för nya tecken är interaktion med teckenspråkiga direkt på olika platser (det årliga deltagandet i Dövas dag-evenemang och andra dövrelaterade aktiviteter), (Riemer Kankkonen et al. 2018). Data som

samlas in via offlinemetoder, t.ex. att interagera med individer direkt, kan ge ett mer kvalitativt tillvägagångssätt för att samla in teckenvarianter. Dessutom ges lexikonteamet en möjlighet att personligen träffa och interagera med dövsamhället, vilket är en viktig aspekt av allt språkdokumentationsarbete.

Alla nya tecken från ovannämnda källor införlivas i lexikonet. Rangordning av fonologiska och lexikala teckenvarianter bygger på antal förekomster i korpusen, lexikonets enkäter och exempelmeningar, och i vissa fall på andra grunder, t.ex. konsulteras teckenspråkiga invånare om tecken för orten de bor på.

Statistik över förekomster presenteras i varje teckenpost, se figur 3, översta bilden under *Förekomster*.

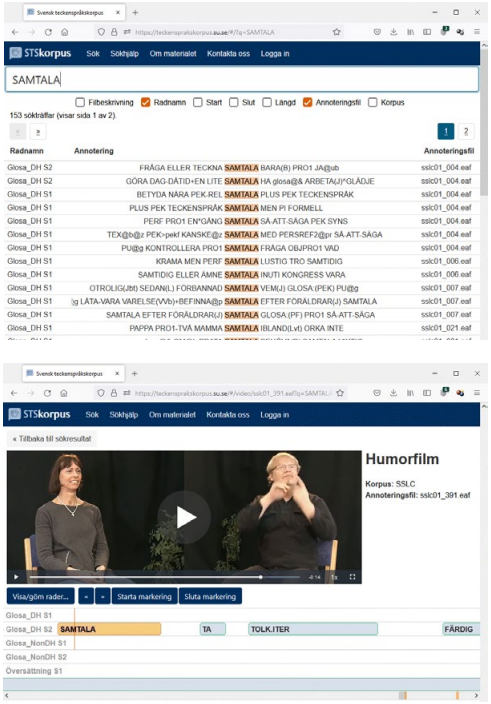
6. Samverkan mellan teckenspråkslexikonet och teckenspråkskorpusen

En korpus är en samling texter, med skrivet eller talat språk, som är transkriberade med ord och uppmärkta med ordens ordklass eller satsdelsfunktion. Nu finns det också en teckenspråkskorpus (Mesch 2023; Mesch & Wallin 2012, 2015; Mesch, Wallin, & Björkstrand 2012; Mesch, Wallin, Nilsson, & Bergman 2012). Den första korpusen med svenskt teckenspråk från samtal mellan två tecknare skapades 2004 i samband med ECHO-projektet (Crasborn et al. 2007). Den andra omfattande teckenspråkskorpusen (SSLC) består av 24 timmars videoinspelningar med samtal, berättelser och presentationer från 42 tecknare, insamlade under åren 2009–2011. (Mesch, Wallin, Nilsson et al. 2012; Mesch 2015). Det sistnämnda arbetet finansierades av Riksbankens Jubileumsfond.

Svensk teckenspråkskorpus innehåller för närvarande cirka 190 000 teckenförekomster. Det inspelade materialet annoterades och transkriberades med hjälp av det multimodala verktyget ELAN (ELAN 2022; Wittenburg et al. 2006). Verktyget används för annotering och för länkning av transkriptioner till digitaliserat video- och audiomaterial. Särskilda annoteringskonventioner har tagits fram och utvecklats genom arbetet med korpusen (Mesch & Wallin 2021) samt bruksanvisning för annotering av teckenspråkstexter (Mesch & Cortes 2021). En del redigerings- och annoteringsarbete kvarstår, eftersom annoteringsarbetet är mycket tidskrävande (Mesch 2023).

Korpusen är ännu mycket liten jämfört med talkorpusar som används för att systematiskt undersöka variation. Korpusens storlek och det faktum att det under insamlingen av data inte förelåg någon explicit lexikal variationsuppgift (se Stamp et al. 2015), innebär i nuläget begränsningar i hur många teckensynonymer eller formvariationer som kan undersökas enbart med användningen av korpusdata.

Samverkan mellan lexikonet och korpusen gäller bl.a. lexikografiska frågor, främst tecken-lemmatisering. Hur ett tecken ska definieras som ett ”kärntecken” i teckenspråkslexikonet är fortfarande oklart, i synnerhet när det kommer till sammansättningar från talat/skrivet språk (mycket vanligt inom ämnesområden som medicin, och i tekniska eller yrkesinriktade ord). Annan viktig fråga rör produktiva tecken (se fotnot 2) som är svåra att beskriva i teckenordböcker, eftersom man då måste fastställa begränsade betydelser hos tecknen, vilket inte stämmer överens med deras användning. Denna teckenkategori finns i korpusen men inte än i lexikonet, och samverkan mellan lexikonet och korpus kan användas i att lösa denna fråga.



FIGUR 7. Skärmbilder från STS-korpus. Överst: sökträffar på tecknet SAMTALA. Underst: ställe i inspelningens tidslinje där en av träffarna förekommer, tillsammans med annoteringar av det tecknade innehållet.

Samverkan mellan Svenskt teckenspråkslexikon och Svensk teckenspråkskorpus har också resulterat i det webbaserade verktyget STS-korpus (Öqvist et al. 2020), som kopplar ihop dessa två resurser. Korpusverktyget hämtar material från korpusen (se figur 7) och används för utökad presentation av hur tecken förekommer i naturlig språkanvändning. STS-korpus inkluderar alla uppmärkta teckenglosor, produktiva tecken (s.k. avbildande tecken[2]) och andra teckenkategorier. Verktyget samverkar med Svenskt teckenspråkslexikon vilket gör teckenexemplen sökbara i båda riktningar, från lexikon eller STS-korpus. När ändringar eller tillägg av nya teckenglosor har gjorts i lexikondatabasen uppdateras de automatiskt i korpusen genom länkade ELAN-filer, vilket underlättar annoterares och forskares arbete. På detta sätt får forskare och lexikon- och korpusanvändare samma information om tecken.

Teckenspråkslexikon och -korpusar är per definition flerspråkiga, i.o.m. att översättning till det omgivande skriftspråket ingår. Dessa resurser är därför särskilt lämpade för andraspråksinlärning (Leeson et al. 2019). Språkutvecklande applikationer, kopplade till Svenskt teckenspråkslexikon, möjliggör inlärare att studera på egen hand, t.ex. TSP Quiz. Detta är ett populärt digitalt verktyg där användare får öva på slumpvist valda tecken från teckenspråkslexikonet. Twitterboten *Ett tecken varje dag* skapades i januari 2018, och har blivit omtyckt av många, [@allatecken skapad av C. Börstell, <https://twitter.com/allatecken>] (nedlagd våren 2023). Den publicerar ett slumpmässigt valt tecken från teckenspråkslexikonet.

7. Avslutning

Att utveckla och underhålla ett teckenspråkslexikon är tids- och resurskrävande. Det är ytterst nödvändigt att hänga med i den tekniska utvecklingen, genom att till exempel uppdatera videofiler till rätt format, och vidta åtgärder för att skydda mot krascher och hackare. Lexikonet behöver också forskning, t.ex. i teckenspråksfonologi, och tekniskt stöd för att hålla hög kvalitet och förbli en användarvänlig resurs.

2 Avbildande tecken saknar grundform. De bildas genom produktiva processer, där den tecknande konstruerar en teckenform som är styrd av den situation som beskrivs. Handformerna väljs i enlighet med språkliga konventioner, medan däremot händernas läge, rörelser och relation till varandra varierar beroende på sammanhanget.

Teckenspråkslexikon behövs för språkdokumentation och för att bidra till att höja teckenspråkets ställning. Lexikonet är även nödvändigt som språkresurs för att tillgodose behoven hos teckenspråkstolkar och döva, teckenspråkslärare och studenter, personer med särskilda behov, samt forskare och andra användare av teckenspråk. Förutom att fungera som språkdokumentation och språkresurser, kan teckenspråksdatabaser användas för framtida utveckling av automatisk översättning (Kopf et al. 2021).

Referenser

Bergman, Brita & Thomas Björkstrand (2015): *Teckentranskription*. FOT-rapport (Forskning om teckenspråk) XXV. Stockholm: Institutionen för lingvistik, Stockholms universitet.

Crasborn, Onno, Johanna Mesch, David Waters, Els van der Kooij, Bencie Woll & Brita Bergman (2007): Sharing sign language data online: Experiences from the ECHO project. *International Journal of Corpus Linguistics*, 12:4, 535–562.

ELAN (Version 6.4) [Computer software] (2022): Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive. <https://archive.mpi.nl/tla/elan>. Hämtat december 2022.

Fenlon, Jordan, Cormier Kearsy & Adam Schembri (2015): Building BSL SignBank: The lemma dilemma revisited. *International Journal of Lexicography*, 28:2, 169–206.

Kopf, Maria, Marc Schulder & Thomas Hanke (2021): *Overview of Datasets for the Sign Languages of Europe*. <https://doi.org/10.25592/UHHFDM.9561>. Hämtat december 2022.

Leeson, Lorraine, Jordan Fenlon, Johanna Mesch, Cermal Grehan & Sarah Sheridan (2019): The uses of corpora in L1 and L2/Ln sign language pedagogy. I: Rosen, Russell S. (red.), *The Routledge Handbook of Sign Language Pedagogy*, 339–352.

McKee, Rachel & Mireille Vale (2017): Sign language lexicography. I: Hanks, Patrick & Gilles-Maurice de Schryver (red.), *International Handbook of Modern Lexis and Lexicography*. Springer Berlin Heidelberg, 1–22.

Mesch, Johanna (2015): *Svensk teckenspråkskorpus – dess tillkomst och uppbyggnad*. FOT-rapport (Forskning om teckenspråk) XXIV. Stock-

holm: Institutionen för lingvistik, Stockholms universitet, 1–25.
Mesch, Johanna (2023): Creating a multifaceted corpus of Swedish Sign Language: Visual, tactile, and L2 signing. I: Wehrmeyer, Ella (red.), *Advances in Sign Language Corpus Linguistics*. Studies in Corpus Linguistics 108. John Benjamins, 242–261. https://doi.org/10.1075/scl.108.09mes
Mesch, Johanna & Elisabet E. Cortes (2021): *Bruksanvisning för annotering av teckenspråkstexter i ELAN* (Version 4). Stockholm: Institutionen för lingvistik, Stockholms universitet.
Mesch, Johanna & Lars Wallin (2012): From meaning to signs and back: Lexicography and the Swedish Sign Language Corpus. I: Crasborn, Onno, Eleni Efthimiou, Evita Fotinea, Thomas Hanke, Jette Kristoffersen & Johanna Mesch (red.), *Proceedings of the {LREC2012} 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon*. European Language Resources Association (ELRA), 123–126.
Mesch, Johanna & Lars Wallin (2015): Gloss annotations in the Swedish Sign Language Corpus. *International Journal of Corpus Linguistics*, 20:1, 103–121.
Mesch, Johanna & Lars Wallin (2021): *Annoteringskonventioner för teckenspråkstexter. Version 8, maj 2021*. Stockholm: Institutionen för lingvistik, Stockholms universitet.
Mesch, Johanna, Lars Wallin & Thomas Björkstrand (2012): Sign Language Resources in Sweden: Dictionary and Corpus. I: Crasborn, Onno, Eleni Efthimiou, Evita Fotinea, Thomas Hanke, Jette Kristoffersen & Johanna Mesch (red.), *Proceedings of the {LREC2012} 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon*. European Language Resources Association (ELRA), 127–130.
Mesch, Johanna, Lars Wallin, Anna-Lena Nilsson & Brita Bergman (2012): Svensk teckenspråkskorpus: Datamängd. Projektet korpus För det svenska teckenspråket 2009–2011 (Version 1). Avdelningen för teckenspråk, Institutionen för lingvistik, Stockholms universitet.
Riemer Kankkonen, Nikolaus, Thomas Björkstrand, Johanna Mesch & Carl Börstell (2018): Crowdsourcing for the Swedish Sign Language Dictionary. I: Bono, Mayumi, Eleni Efthimiou, Stavroula-Evita Fotinea, Thomas Hanke, Julie Hochgesang, Jette Kristoffersen, Johanna

Mesch & Yutaka Osugi (red.), *Proceedings of the {LREC2018} 8th Workshop on the Representation and Processing of Sign Languages: Involving the Language Community*. European Language Resources Association (ELRA), 171–174.

Schermer, Gertrude M. (2006): Sign language: Lexicography. I: Brown, Keith (red.), *Encyclopedia of Language & Linguistics*. Elsevier, 321–324.

Stamp, Rose, Adam Schembri, Jordan Fenlon & Ramas Rentelis (2015): Sociolinguistic variation, language change and contact in the British Sign Language (BSL) lexikon. *Sign Language & Linguistics* 18:1, 158–166.

Svenskt teckenspråkslexikon (1997): Sveriges Dövas Riksförbund.

Wittenburg, Peter, Hennie Brugman, Albert Russel, Alex Klassmann & Han Sloetjes (2006): ELAN: a professional framework for multimodality research. I: *Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation*. European Language Resources Association (ELRA).

Öqvist, Zrajm, Nikolaus Riemer Kankkonen & Johanna Mesch (2020): STS-korpus : A sign language web corpus tool for teaching and public use. I: Efthimiou, Eleni, Evita Fotinea, Thomas Hanke, Julie Hochgesang, Jette Kristoffersen & Johanna Mesch (red.), *Proceedings of the {LREC2020} 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*. European Language Resources Association (ELRA), 177–180.

Österberg, Oskar (1916): *Teckenspråket*. P. Alfr. Persons förlag.

Överförda betydelser, semantiska utvidgningar och andra oegentligheter. En undersökning av fem definitionsformler för semantisk förändring i SAOB:s definitionstext

Pär Nilsson

In my Ph. D. project, I study semantic change as it is described in the Swedish Academy dictionary (SAOB). The starting point is five semantic labels that are established and highly frequent in the dictionary definitions, namely: *bildlig* ('figurative'), *oegentlig* ('un-actual/non-factual'), *överförd* ('transferred'), *utvidgad* ('extended') and *allmännare användning* ('more general use'). These labels are examined with different methods and from different perspectives.

The question posed is how the labels are used in the SAOB, and how systematic and consequent the use is. The aim is to find out what the labels in question mean in practice: What is a figurative use in the dictionary? What is the difference between an extended and a more general use, and what, more precisely, does it mean in the dictionary that a meaning could be un-actual?

The purpose of the survey is to assess how relevant and consistent the analysis and description methods applied in the SAOB are today, from a theoretical perspective. To what extent is the description of the mechanisms behind semantic change compatible with modern semantic theory and in particular the ideals of cognitive semantics? From the opposite perspective, the purpose is to find out what the definitions of SAOB can teach us about semantic change and the meaning-changing mechanisms.

In this article, the project is described in more detail, and some preliminary results are presented.

NYCKELORD: lexikografi, semantisk förändring, kognitiv semantik, konceptuell metaforteori

1. Inledning

I mitt doktorandprojekt, som bedrivs på Språk- och litteraturcentrum vid Lunds universitet, studeras betydelseutveckling såsom den kommer till uttryck i *Svenska Akademiens ordbok* (SAOB). Undersökningen tar sin utgångspunkt i fem s.k. definitionsformler som förekommer frekvent

i ordbokens definitionsspråk, närmare bestämt *bildl.*, *överförd*, *oeg.*, *utvidgad* samt *allmännare anv*. I föreliggande artikel sammanfattas och beskrivs projektet lite närmare och några preliminära resultat presenteras. Dessa resultat avser en av avhandlingens delundersökningar, nämligen delstudie 1 (se vidare avsnitt 1.3 nedan) som tar sikte på vilka semantiska processer som finns representerade i undersökningsmaterialet.

1.1. Syfte

Syftet med projektet är tvådelat. För det första vill jag ta reda på vad det går att lära sig om semantisk förändring genom att studera ordbokens formelförsedda definitioner. För det andra – och ur motsatt perspektiv – är syftet att undersöka hur semantikteoretiska diskussioner som förs i litteraturen kan föra ordbokens arbete framåt. Ur detta senare perspektiv analyseras de aktuella definitionsformlerna, och SAOB:s pragmatiskt utvecklade beskrivningsmetod, med utgångspunkt i semantisk teori, inte minst med grundbegrepp hämtade från den kognitiva semantiken. Med tanke på att den första upplagan av ordboken inom kort kommer att färdigställas (sista halvbandet publiceras under 2023) och att en omfattande revidering står för dörren, är målsättningen att undersökningens resultat ska få direkt tillämpning i redaktionens kommande arbete.

1.2. Definitionsformlerna – urval, bakgrund och några exempel

Innan frågeställningen specificeras närmare och metod och material beskrivs bör några ord sägas om de aktuella definitionsformler som studeras.

SAOB:s metaspråk totalt sett är nämligen i hög grad formelartat, och det finns många andra termer i ordboken, förutom de aktuella fem formlerna, som tar sikte på semantiska och/eller pragmatiska och grammatiska förhållanden, men som inte studeras i projektet. Gemensamt för dem som undersöks är att de beskriver betydelse*utveckling*, och det är de fem mest frekventa formlerna i ordboken som har valts ut. Vidare har krävts att formlerna ska beteckna *denotativ* och *semasiologisk* förändring i Geeraerts' (2015) bemärkelse. Begreppet denotativ avser här referentiell betydelse (vilket kan kontrasteras mot exempelvis rent emotiv betydelse). Semasiologisk innebär ett perspektiv som utgår ifrån det individu-

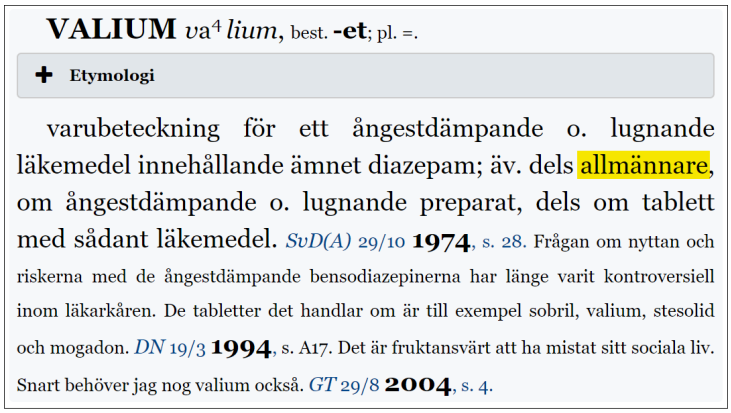
ella ordet och de olika begrepp som detta betecknar (i motsats till termen *onomasiologisk*, som i stället fokuserar på det abstrakta begreppet och vilka ord som kan användas för att uttrycka detta).

Den förkortade formeln *bildl. anv.*, utläst: *bildlig användning*, förekommer drygt 40 000 gånger i ordbokens spalter mellan bokstäverna A och Å (dvs. så långt fram i alfabetet som ordboksarbetet hunnit i skrivande stund). Etiketten *oeg. anv. (oegentlig användning)* påträffas ca 11 000 gånger, *utvidgad* ca 10 000, *allmännare* knappt 8 000 och *överförd anv.* totalt ca 5 000 gånger.

Dessa formler är traditionella i den meningen att det inte är SAOB-redaktionen själv som har myntat termerna. De förekommer i flera andra ordböcker och ofta verkar SAOB-redaktionen ha hämtat inspiration från dem. Som exempel kan nämnas formuleringar som *transferred* och *extended use* i *Oxford English Dictionary* (OED) och *bildlich* och *uneigentlich* i *Deutsches Wörterbuch von Jacob Grimm und Wilhelm Grimm* (DWB). Begrepp som bildlighet och överföring diskuteras ibland också i den semantiska litteraturen.

Men vad dessa olika termer och begrepp egentligen betecknar, mer exakt, är ofta vagt definierat. Det är inte heller glasklart formulerat i SAOB:s interna manualer, trots att de omnämns där. Mot denna bakgrund förefaller en närstudie av de fem definitionsformlerna således motiverad: hur ska läsaren förstå dessa etiketter som så ofta dyker upp i ordbokens definitioner?

För att illustrera hur formlerna kommer till uttryck i praktiken i SAOB följer här nedan ett par exempel hämtade ur definitionstexten. I exemplen har de aktuella formlerna gulmarkerats för att framhävas tydligare för läsaren (denna princip tillämpas sedan i resten av artikeln). Notera att formlerna ofta kombineras med varandra – i exemplet SMED beskrivs den härledda betydelsen som oegentlig eller bildlig.

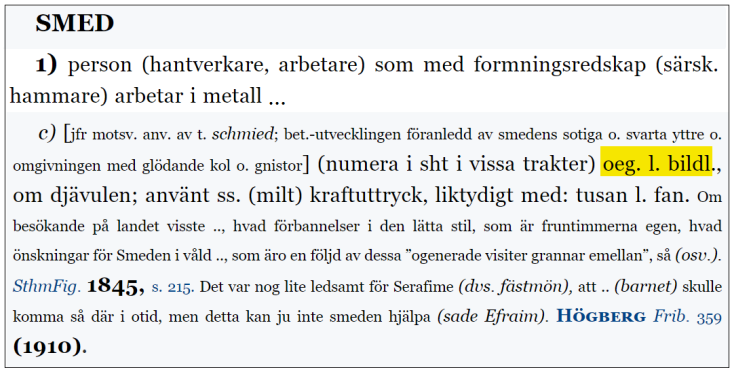


VALIUM *va*[4] *lium*, best. **-et**; pl. =.

✚ **Etymologi**

varubeteckning för ett ångestdämpande o. lugnande läkemedel innehållande ämnet diazepam; äv. dels allmännare, om ångestdämpande o. lugnande preparat, dels om tablett med sådant läkemedel. *SvD(A)* 29/10 **1974**, s. 28. Frågan om nyttan och riskerna med de ångestdämpande bensodiazepinerna har länge varit kontroversiell inom läkarkåren. De tabletter det handlar om är till exempel sobril, valium, stesolid och mogadon. *DN* 19/3 **1994**, s. A17. Det är fruktansvärt att ha mistat sitt sociala liv. Snart behöver jag nog valium också. *GT* 29/8 **2004**, s. 4.

FIGUR 1. Allmännare användning, exemplet VALIUM.



SMED

1) person (hantverkare, arbetare) som med formningsredskap (särsk. hammare) arbetar i metall ...

c) [jfr motsv. anv. av t. *schmied*; bet.-utvecklingen föranledd av smedens sotiga o. svarta yttre o. omgivningen med glödande kol o. gnistor] (numera i sht i vissa trakter) oeg. l. bildl., om djävulen; använt ss. (milt) kraftuttryck, liktydigt med: tusan l. fan. Om besökande på landet visste .., hvad förbannelser i den lätta stil, som är fruntimmerna egen, hvad önskningar för Smeden i våld .., som äro en följd av dessa ”ogenerade visiter grannar emellan”, så *(osv.)*. *SthmFig.* **1845**, s. 215. Det var nog lite ledsamt för Serafime *(dvs. fästmön)*, att .. *(barnet)* skulle komma så där i otid, men detta kan ju inte smeden hjälpa *(sade Efraim)*. **HÖGBERG** *Frib.* 359 **(1910)**.

FIGUR 2. Oegentlig eller bildlig användning, exemplet SMED.

I exemplet VALIUM har den semantiska extensionen hos namnet på ett läkemedel med en viss verksam substans vidgats, så att det kan användas om (liknande) lugnande medel i allmänhet. Den semantiska förändringen etiketteras som en *allmännare* användning. Vid det andra exemplet (SMED) tänks djävulen metaforiskt som en smed. Denna förändring har fått etiketten *oeg. l. bildl.*

1.3. Frågeställning

Frågan som ställs i avhandlingen är för det första *vad* de fem aktuella definitionsformlerna innebär i praktiken, rent semantiskt. Med utgångspunkt i modern semantisk teori undersöks vilka semantiska processer som är involverade vid de olika formlerna.

För det andras studeras *hur* de språkliga användningar som etiketterats med en formel beskrivs i ordboken. För att besvara frågorna analyseras därför både ordbokens beskrivningar och metaspråk och också den språkanvändning och de språkprov som ordboken beskriver i sina spalter.

Fem olika delstudier genomförs, vilka var och en representerar en särskild forskningsfråga. Som nämnts ovan studeras för det första vilken semantisk process det är frågan om (delstudie 1) – motsvarar exempelvis en bildlig användning alltid en metafor? För det andra undersöks i vilken utsträckning en formelförsedd användning innebär mappning mellan konkreta och abstrakta begrepp (delstudie 2). Frågan grundar sig i tanken att metaforik inom den kognitiva semantiken typiskt beskrivs just som projicering från konkret till abstrakt domän (se t.ex. Kövecses 2017) – gäller detta även i SAOB?

Som en tredje delstudie undersöks var i ordbokens mikrostruktur som definitionsformeln påträffas: står den ihop med grundbetydelsen, eller på en separat och mer självständig position? Delstudien utgör ett försök att ta mått på graden av autonomi hos de härledda betydelserna, och att undersöka hur lika de aktuella begreppsdomänerna och käll- och målbegreppen är.

I den fjärde delstudien noteras i vilken utsträckning de aktuella formlerna modifieras eller kombineras med andra formler (som vid exemplet SMED ovan, där den härledda betydelsen etiketterades som *oeg. l. bildl.*). I den kognitivistiska litteraturen beskrivs gränsen mellan metafor och metonymi som en glidande skala, och ett språkligt uttryck kan vara mer eller mindre metaforiskt eller metonymiskt. Den aktuella delstudien blir ett sätt att studera gråzoner mellan olika semantiska processer. Vilka formler kombineras och hur, och på vilket sätt skiljer sig de dubbeletiketterade användningarna från de som etiketterats med en enda formel?

I den femte delstudien, till sist, undersöks vilka ordklasser som de etiketterade betydelseutvecklingarna representerar. Metaforik och metonymi har ibland beskrivits som substantivfenomen (jfr t.ex. diskussionen i Croft 1993), och exempelvis adverb och adjektiv omnämns ibland som bortglömda ordklasser i fråga om semantiska analyser. I delundersökningen studeras alltså i vilken utsträckning de olika definitionsformlerna är låsta till vissa ordklasser.

Undersökningen avser att mynna ut i en diskussion av hur de olika aspekter som studeras tillsammans målar upp en bild av vad en etiketterad betydelse i SAOB är för något.

Som nämndes ovan riktas i föreliggande artikel fokus dock i huvudsak mot en av de fem aspekter som analyseras. I det följande diskuteras semantiska processer. Den aktuella forskningsfrågan att besvara gäller alltså vad det är rent semantiskt som formlerna *bildl.*, *oeg.*, *utvidgad*, *allmännare* och *överförd anv.* betecknar.

2. Teoretiska grundbegrepp

Som redan nämnts lutar sig undersökningen mot semantisk teori och i synnerhet kognitiv semantik och konceptuell metaforteori. Utgångspunkten vid den semantiska analysen är framför allt de tre semantiska processerna *metafor*, *metonymi* och *generalisering* (men äv. t.ex. *ellips* och *liknelse*), och analysen blir ett försök till matchning: i vilken utsträckning motsvaras exempelvis en användning som i SAOB betecknats som *bildl.* av en metafor, och en *allmännare anv.* av en generalisering?

Analysen utgår ifrån Lakoff & Johnsons (1980) definition av metafor och metonymi: metaforer ska inte bara betraktas som ett fenomen i språket, utan som något som är relaterat till vårt bakomliggande sätt att tänka och agera. Metaforik innebär mappning mellan två olika begreppsdomäner; och med en domän avses en strukturerad helhet, en sammanhängande organisation av erfarenheter.

Metonymi innebär mappning inom en och samma domän, eller ett och samma domänkomplex (jfr Croft 1993), och ett samtidigt skifte av referens.

Begreppet *generalisering* är av olika anledningar inte lika frekvent diskuterat inom konceptuell metaforteori. Men vad som brukar nämnas i diskussionen är att det går att ifrågasätta ifall det är en egen process eller bara resultatet av en annan process: om generalisering ska förstås endast som att betydelsen eller betydelsekategorin utvidgas så blir ju metafor en undertyp av generalisering (jfr Bybee et al. 1994). I avhandlingen betraktas dock (i enlighet med Koch 2016) generalisering och semantisk utvidgning som två separata företeelser. I den mån domänmappning kan förklaras som ett gradfenomen bör generalisering befinna sig på samma skala som metafor – också här handlar det om analogi mellan ett käll- och ett målbegrepp, men för generaliseringens del är det frågan om mappning inom en och samma domän (jfr exemplet VALIUM i avsnitt 1.2 ovan där källa och mål i båda fallen hör till domänen medicin). Skillnaden

mot metonymi ligger i att den senare processen alltså kräver ett samtidigt skifte av referens. Både generalisering och metafor leder dock till *semantisk utvidgning*.

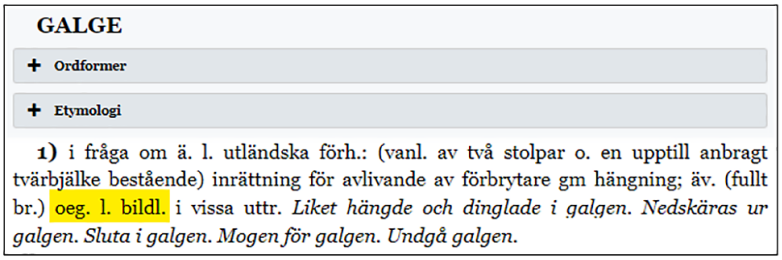
Som nämndes i föregående avsnitt talar man inom den kognitiva semantiken ofta om att de olika semantiska processerna ska betraktas som gradfenomen. En metafor kan ha större eller mindre metaforisk styrka, beroende på i vilken utsträckning föreställningar om källdomänen aktiveras vid mappningen (jfr t.ex. Svanlund 2007, 2009). Ur detta perspektiv blir skillnaden mellan en metafor med låg styrka och en generalisering otydlig och gränsen processerna emellan svår att dra.

3. Material och metod

Det material som undersöks utgörs av 300 olika förekomster vardera (dvs. 1 500 förekomster totalt) av varje formel i den digitala versionen av SAOB. För att få en spridning av materialet, inom alfabetet och över tid i ordbokens utgivningshistoria, har varje omgång om 300 formelförekomster delats i fyra. 75 förekomster vardera har sökts fram inom var och en av de fyra olika perioder som Larsson (2014) diskuterar. Perioderna representerar olika normer sinsemellan med avseende på ordboksartiklarnas omfång och definitionernas utförlighet. Period 1 sträcker sig från 1890–1920, Period 2 från 1920–1960, Period 3 1960–2002 och Period 4: 2002–.

Efter att undersökningsmaterialet har excerperats har varje artikel där den aktuella formeln förekommer närstuderats. En analysmall har tillämpats som tar sikte på den forskningsfråga som beskrevs i avsnitt 1.3 ovan. Förutom hur själva formeln i sig är utformad har även definitionstexten i sin helhet analyserats och de autentiska språkprov som belägger och illustrerar den aktuella användningen.

I praktiken har det då visat sig att en och samma formel ibland beläggs med språkprov som representerar mer än en semantisk process. Så är t.ex. fallet vid exemplet GALGE här nedan. Vid detta exempel går den semantiska processen att tolka som inbegripande både metafor och metonymi.


GALGE

+ Ordformer

+ Etymologi

1) i fråga om ä. l. utländska förh.: (vanl. av två stolpar o. en upptill anbragt tvärbjälke bestående) inrättning för avlivande av förbrytare gm hängning; äv. (fullt br.) oeg. l. bildl. i vissa uttr. *Liket hängde och dinglade i galgen. Nedskäras ur galgen. Sluta i galgen. Mogen för galgen. Undgå galgen.*

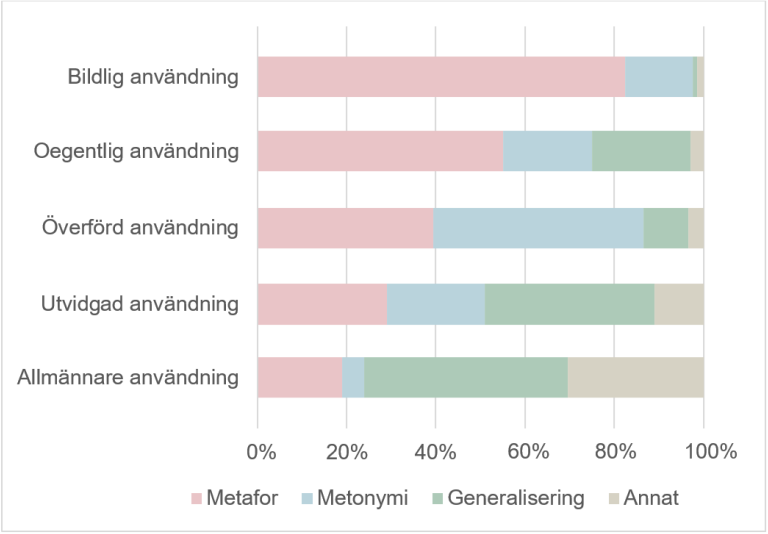
FIGUR 3. Etiketterad användning i SAOB som kan tolkas som metafor och/eller metonymi, exemplet GALGE.

Vid ett uttryck som (*någon kommer att*) *sluta i galgen* står ordet *galge* metonymiskt för avrättning. Innebörden bör alltså tolkas som att någon kommer att dö. Uttrycket kan dock även tolkas metaforiskt (hyperboliskt) som en allmännare beteckning för att det kommer att sluta illa för någon. Vid dylika fall i materialet har mer än en semantisk process noterats för varje formelförekomst.

4. Några preliminära resultat – semantiska processer

I det följande presenteras resultat från delstudie 1 som har att göra med vilka semantiska processer som finns representerade i materialet – vad står de olika formlerna för rent innehållsligt, och i vilken utsträckning motsvarar de metafor, metonymi och generalisering?

I nedanstående diagram presenteras en sammanställning av de övergripande mönster som kommer till uttryck i undersökningsmaterialet. I figuren har det totala antalet instantieringar av varje formel räknats samman. Det innebär att de fall (likt GALGE i figur 3 ovan) där en formel representeras av mer än en semantisk förändringsprocess har resulterat i två olika instantieringar av denna formel. Det föreligger således en skillnad mellan antalet formelförekomster (300 förekomster per formel) och antalet analysenheter (varierande per formel). För jämförbarhetens skull är det andelen, och inte antalet, av varje semantisk process som redovisas i diagrammet.



FIGUR 4. Sammanställning semantiska processer, instantieringar per formel.

Vad som går att utläsa av diagrammet är att processtyp i materialet verkar vara en glidande skala mellan de olika formlerna, och att ingen formel exklusivt representerar en enda process. Samtliga formler involverar ett visst mått av olika förändringsprocesser.

Men samtidigt verkar varje formel ha en ganska tydlig egen semantisk profil. Bildlig användning utgörs t.ex. i huvudsak av metaforer, men även en del metonymiska användningar påträffas hos denna formel. Motsatt förhållande råder för de överförda användningarna, i mitten av figuren. Där dominerar i stället metonymi, men även en hel del metaforik förekommer.

De oegentliga användningarna spretar en del, men utgörs övervägande av metaforer. Formeln betecknar dock även i tämligen hög utsträckning generalisering.

De utvidgade och allmännare användningarna är ganska jämnt fördelade med avseende på semantisk process. De domineras båda av generaliseringar, men involverar även en del andra processer, både metaforer och metonymier.

Anledningen till att fördelningen ser ut som den gör och att varje formel trots tydlig egen prägel involverar flera olika förändringsprocesser går förstås inte att utläsa av diagrammets sammanställning, och fortsatt analysarbete återstår innan avhandlingens delundersökning är färdigställd.

Men redan nu kan några iakttagelser från analysen av materialet göras som åtminstone till viss del förklarar skillnaderna formlerna emellan.

4.1. Formlerna *bildl.* och *överförd anv.*

Som redan nämnts betecknar en formel ibland mer än en semantisk process, och detta är ofta fallet för de bildliga användningar som betecknar metonymi. Majoriteten av dessa utgörs av typen som exemplifierades i figur 3 där metafor och metonymi samförekommer – ett slag av *metaftonymi* i Goossens (1990) termer.

För de överförda användningarna gäller att de i den första, äldre, halvan av undersökningsmaterialet (dvs. under period 1–2) i störst utsträckning användes som beteckning för metaforer. Under andra halvan av materialet (period 3–4) betecknar formeln i stället i huvudsak metonymier, och i synnerhet den typ som förekommer i figur 5 här nedan, där en personbetecknande egenskap projiceras på något sakligt (eller vice versa). I språkprovet från 1898 (”talangfulla dukar”) har talangfullheten metonymiskt ”överförts” från konstnären till verket.

TALANG-FULL. om person: full av talang, som har (stor) talang l. (många) talanger; äv. i överförd anv., om ngt sakligt: som vittnar om l. kännetecknas av talang; jfr **-rik.** *JournLTh.* **1812**, nr 86, s. 1. *GHT* **1898**, nr 265, s. 2 **(**: talangfulla dukar**)**. *(Den)* kraftige, dugande och talangfulle .. kapten Lars Wilhelm Kylberg .. blev fader till en stor barnskara. *Fatab.* **1952**, s. 125.

FIGUR 5. Överförd användning som utgörs av metonymi; exemplet TALANG-FULL.

Skiftet av innebörd hos formeln (från metafor i tidigare delar av ordboken, till metonymi i senare delar) grundar sig i en förändring av de redaktionella riktlinjerna. I Hans Jonssons interna handbok (1993) påpekas att formeln tidigare har använts utan klar åtskillnad från *bildl.*, och att den i stället bör användas för den här ovan exemplifierade metonymitypen (i handboken ges bl.a. exemplet: *ett vänligt brev*). Uppmaningen verkar således ha fått tydligt genomslag.

4.2. Formeln *oeg. anv.*

Om de oegentliga användningarna kan två iakttagelser göras. För det första verkar de metaforer som försetts med formeln skilja sig lite åt från dem som etiketterats som bildliga (eller överförda). Vid en närmare granskning visar sig flera av de ”oegentliga” metaforerna uppvisa en jämförelsevis låg metaforisk styrka. De utgörs nämligen ofta av mappningar mellan förhållandevis lika domäner, och inte sällan är det frågan om käll- och måldomäner med likartad abstraktionsgrad. Dessa kan ställas i kontrast till den standarddefinition av metafor som Kövecses (2017) ger, enligt vilken metaforisk mappning typiskt sker mellan konkret källdomän och abstrakt måldomän.

Figur 6 nedan utgör ett typiskt exempel på sådan användning i materialet. I den oegentliga användningen av ordet *taliban* projiceras fundamentalistisk islamism metaforiskt på intolerans och sträng personlighet. I både käll- och måldomän handlar det om personer.

TALIBAN

(i sht i fråga om förh. i Afghanistan) om aktivist inom ultraortodox o. militant sunnimuslimsk studentrörelse (vars mål är inrättandet av en strikt islamistisk stat); ss. förled i ssgr äv. om rörelsen (se **RÖRELSE 8**).

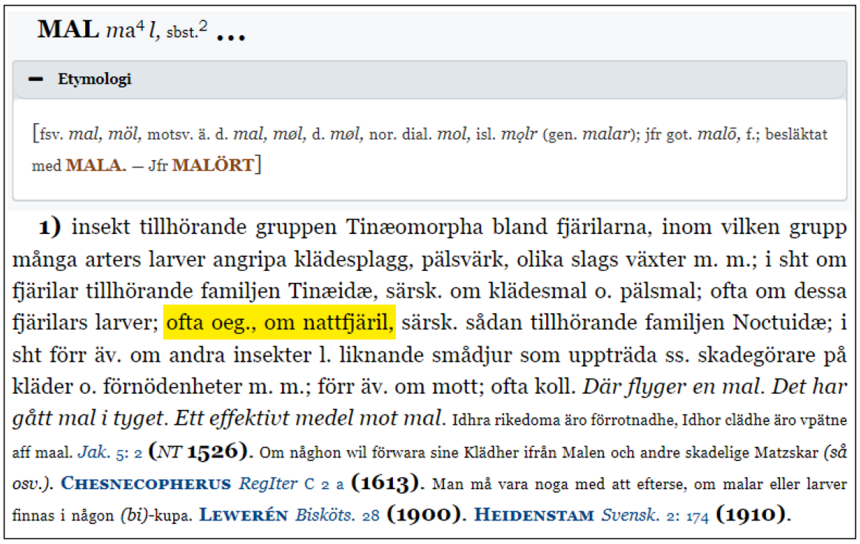
...

särsk. (nedsättande) i oeg. anv., om person ss. representant l. förespråkare l. dyl. för extrem konservatism (l. stränghet) l. intolerans l. förbud l. förtryck (i sht av kvinnor) o. d. *DN* 19/10 **1996**, s. B2. Talibanerna ropade på obligatoriska alkohollås i alla fordon .. samt fler hastighetskontroller. *Därs.* 2/2 **2002**, s. E2. Läste att Gudrun Schyman anmält sig till Tjejvasan! Undrar om anledningen att hon inte valde Stora Vasaloppet är att hon inte vill omge sig med 10 000 talibaner? *KvällsP* 19/2 **2002**, s. 17.

FIGUR 6. Oegentlig användning som utgörs av metafor med låg metaforisk styrka; exemplet TALIBAN.

För det andra är de generaliseringar som etiketterats som oegentliga ofta av ett särskilt slag. Det rör sig nämligen i stor utsträckning om betydelser som inbegriper en normkonflikt mellan allmänspråket och ett striktare fackspråk, som vid exemplet MAL här nedan. I detta fall har betydelsekategorin vidgats på ett icke-metaforiskt vis, men användningen uppfattas i egentlig mening vara felaktig: i fråga om nattfjärilar före-

kommer inget egentligt ”malande” (nattfjärilar angriper vanligen inte klädesplagg).



MAL *ma*[4] *l*, sbst.[2] ...

– Etymologi

[fsv. *mal, möl,* motsv. ä. d. *mal, møl,* d. *møl,* nor. dial. *mol,* isl. *mǫlr* (gen. *malar*); jfr got. *malō,* f.; besläktat med **MALA.** – Jfr **MALÖRT**]

1) insekt tillhörande gruppen Tinæomorpha bland fjärilarna, inom vilken grupp många arters larver angripa klädesplagg, pälsvärk, olika slags växter m. m.; i sht om fjärilar tillhörande familjen Tinæidæ, särsk. om klädesmal o. pälsmal; ofta om dessa fjärilars larver; ofta oeg., om nattfjäril, särsk. sådan tillhörande familjen Noctuidæ; i sht förr äv. om andra insekter l. liknande smådjur som uppträda ss. skadegörare på kläder o. förnödenheter m. m.; förr äv. om mott; ofta koll. *Där flyger en mal. Det har gått mal i tyget. Ett effektivt medel mot mal.* Idhra rikedoma äro förrotnadhe, Idhor clädhe äro vpätne aff maal. *Jak.* 5: 2 **(*NT* 1526)**. Om någhon wil förwara sine Klädher ifrån Malen och andre skadelige Matzskar *(så osv.)*. **CHESNECOPHERUS** *RegIter* C 2 a **(1613)**. Man må vara noga med att efterse, om malar eller larver finnas i någon *(bi)*-kupa. **LEWERÉN** *Bisköts.* 28 **(1900)**. **HEIDENSTAM** *Svensk.* 2: 174 **(1910)**.

FIGUR 7. Oegentlig användning som utgörs av generalisering som inbegriper normkonflikt; exemplet MAL.

4.3. Formlerna *utvidgad* och *allmännare anv.*

Av diagrammet i figur 4 ovan kan alltså utläsas att formlerna *utvidgad* och *allmännare anv.* involverar flera metaforiska och metonymiska processer. Att dessa formler inte enbart betecknar generaliseringar, vilket man intuitivt skulle kunna förvänta sig, hänger till stor del ihop med att de också används övergripande, om resultatet av den semantiska processen, och inte enbart såsom en etikett för processen i sig. Detta förhållande blir tydligt i sådana fall som ALLMÄNNING, här nedan, där formeln *allmännare* explicit övergriper *bildl. anv.* (Redaktören verkar alltså här mena att en bildlig användning kan ses som ett slag av en allmännare användning.)

ALLMÄNNING …

2) allmän, gemensam egendom; i sht om ett område, hvartill egande- (l. nyttjande-)rätten tillkommer kronan l. ett större l. mindre kommunalt samfund l. grannarna gemensamt.

a) mark (särsk. skogs- o. betesmark), som utgör gemensam egendom.

…

3) [slutande sig till **2,** särsk. **2** *a*] i allmännare, vanl. bildl. anv. *(Pindarus')* bilder äro icke af det vanliga och förslitna slaget, som en hvar med måttlig beläsenhet och talang kan hämta från den poetiska allmänningen; de äro tvärt om utmärkta genom sin nyhet, kraft och dristighet. **TEGNÉR** 3: 512 **(1816).** Ännu var ej predikan en allmänning, som tillhörde ingen, derföre att den tillhörde alla. **WIESELGREN** *Sv:s sköna litt.* 1: 258 **(1833).** Man betar vidt och bredt på den konstitutionela allmänningen. **TEGNÉR** 3: 381 **(1836;** *om politiskt kannstöperi i sällskapslifvet***).** En svart sidenkappa .. (en allmänning inom Frankska familjen). **BREMER** *Hem.* 1: 168 **(1839).** Inom hednaverldens stora öcken ligger redan mången grönskande inhägnad, vunnen från den vilda allmänningen. **MELIN** *Pred.* 3: 45 **(1852).** *(De tekniska områdena)* äro ett slags allmänning för den europeiska odlingen. **V. RYDBERG** i *Sv. Tidskr.* **1873,** s. 515.

FIGUR 8. Allmännare användning såsom en etikett som är överordnad bildlig användning; exemplet ALLMÄNNING.

5. Sammanfattning och slutsatser

En undersökning av betydelseutveckling såsom den kommer till uttryck SAOB har genomförts med utgångspunkt i fem lexikografiska definitionsformler (och tre semantiska grundbegrepp i den teoretiska litteraturen). Resultaten visar att beskrivningen i ordboken är relativt konsekvent genomförd; varje formel har en ganska tydlig egen semantisk profil, men det är inte på något sätt frågan om ett 1:1-förhållande mellan semantisk process och formel. I stället är det en glidande skala, och varje undersökt definitionsformel betecknar i viss utsträckning samtliga undersökta betydelseutvecklingsmekanismer.

Detta senare förhållande innebär i vissa delar otydlighet för ordboksanvändaren. Formeln *oeg*. avser ibland metaforer och ibland generaliseringar, *överförd anv*. betecknade tidigare i huvudsak metaforer, men numera vanligen metonymier, och formlerna *utvidgad* och *allmännare anv*. verkar ibland ta sikte på semantisk process och ibland på resultatet av en semantisk process. Bruket av definitionsformler i SAOB går här att strama åt. Med tanke på överlappningen mellan några av formlerna går det även att fråga sig om samtliga formler behövs. Formeln *oeg*. exempelvis, som företrädesvis betecknar semantiska förändringar och förhållan-

den som kan uttryckas och beskrivas med andra formler, borde i en reviderad upplaga av ordboken kunna mönstras ut och ersättas av *bildl.* och *utvidgad anv.*

Men resultatet sätter också fingret på att gränsen mellan de olika semantiska processerna är oskarp och att det förekommer gråzoner dem emellan. I det undersökta materialet är ofta mer än en semantisk process involverad i ett och samma betydelseutvecklingsförlopp, t.ex. som vid det metaftonymiska exemplet GALGE i figur 3 ovan. Tendensen i SAOB att markera sådan användning med dubbla definitionsformler (i exemplet ovan var etiketten *oeg. l. bildl.*) kan göra ordboksanvändaren uppmärksam på detta förhållande.

På det hela taget utgör ordbokens bruk av definitionsformler en ganska nyanserad semantisk beskrivning. Mycket tid har lagts ned i det redaktionella arbetet på djupgående analyser och komplexa definitioner – i den andra upplagan av SAOB gäller det nu att göra beskrivningarna lite mer tillgängliga för läsarna.

Litteratur

Ordböcker

DWB = *Deutsches Wörterbuch, von Jacob Grimm und Wilhelm Grimm*, hrsg. von der Deutschen Akademie der Wissenschaften zu Berlin, Band 1–16 + Quellen-verzeichnis, Leipzig 1852–1971.

OED = *The Oxford English Dictionary*, second edition, vol. I–XX (of the corrected re-issue (1933) with an Introduction, Supplement, and Bibliography of A New English Dictionary on Historical Principles, vol. I–X, ed. by James A. H. Murray et al. (1884–1928) and Supplement, vol. I–IV, ed. by R. W. Burchfield (1972–1986)), Oxford 1989.

SAOB = *Ordbok över svenska språket, utgiven av Svenska Akademien*, hittills: bd 1–38, Lund 1893–. Digitalt tillgänglig på: <saob.se>. Hämtat 2022-09-07.

Övrig litteratur

Bybee, Joan L., Perkins, Revere D. & Pagliuca, William 1994. *Evolution of grammar: tense, aspect, and modality in the languages of the world*. Chicago: Univ. of Chicago Press.

Croft, William 1993. The role of domains in the interpretation of metaphors and metonymies. *Cognitive linguistics* 4, 335–370.

Geeraerts, Dirk 2015. How words and vocabularies change. I: Taylor, John R. (ed.), *The Oxford Handbook of the Word*. Oxford: Oxford University Press, 416–430.

Goossens, Louis 1990. Metaphtonymy: the interaction of metaphor and metonymy in expressions for linguistic action. *Cognitive Linguistics* 1–3, 323–340.

Jonsson, Hans 1993. *Handbok för redigeringen av Svenska Akademiens ordbok*. Lund: Svenska Akademiens ordboksredaktion.

Koch, Peter 2016. Meaning change and semantic shifts. I: Juvonen, Päivi & Maria Koptjevskaja-Tamm (eds.), *The lexical typology of semantic shifts*. Berlin: De Gruyter Mouton.

Kövecses, Zoltán 2017. Conceptual metaphor theory. I: Semino, Elena & Zsófia Demjén (eds.), *The Routledge handbook of metaphor and language*. Milton Park, Abingdon, Oxon: Routledge, 13–27.

Lakoff, George & Johnson, Mark 1980. *Metaphors We Live By*. Chicago/London: The University of Chicago Press.

Larsson, Lennart 2014. En ”mer l. mindre stor” stor ordbok – om variationerna i SAOB:s omfång och ambitionsnivå. *LexicoNordica* 21, 61–80.

Svanlund, Jan 2007. Metaphor and convention. *Cognitive Lingustics* 18:1, 47–89.

Svanlund, Jan 2009. *Lexikal etablering: en korpusundersökning av hur nya sammansättningar konventionaliseras och får sin betydelse*. Stockholm: Acta universitatis Stockholmiensis.

Hvordan holde tritt med tiden i en historisk ordbok som også beskriver samtiden?

Carina Nilstun

The Norwegian Academy Dictionary (NAOB) is a historical dictionary, describing the variety bokmål/riksmål from early 1800 until today. It is also a literary dictionary, with entries built upon the actual language use in texts. Furthermore, NAOB is a contemporary dictionary. The entry head reflects today's spelling, inflection and pronunciation, and the editorial examples belong to the present. NAOB is in its origin a reading comprehension dictionary but can today also serve as a production dictionary. NAOB is typologically wide, and to fully be a contemporary dictionary it must receive continuous refills of current material on every level: words, senses and fixed expressions, as well as quotations. Quotations from the recent years should be rich in quantity, given that the base consists of more than 200 000 entries and more than 300 000 quotations. Publications from after year 2000 are not publicly available in Norway. But these 22 years are of immense importance for our contemporary dictionary. In this paper I will show how the lexicographers of NAOB work their way around this.

NØKKELORD: praktisk leksikografi, litterær ordbok, leksikografisk metode

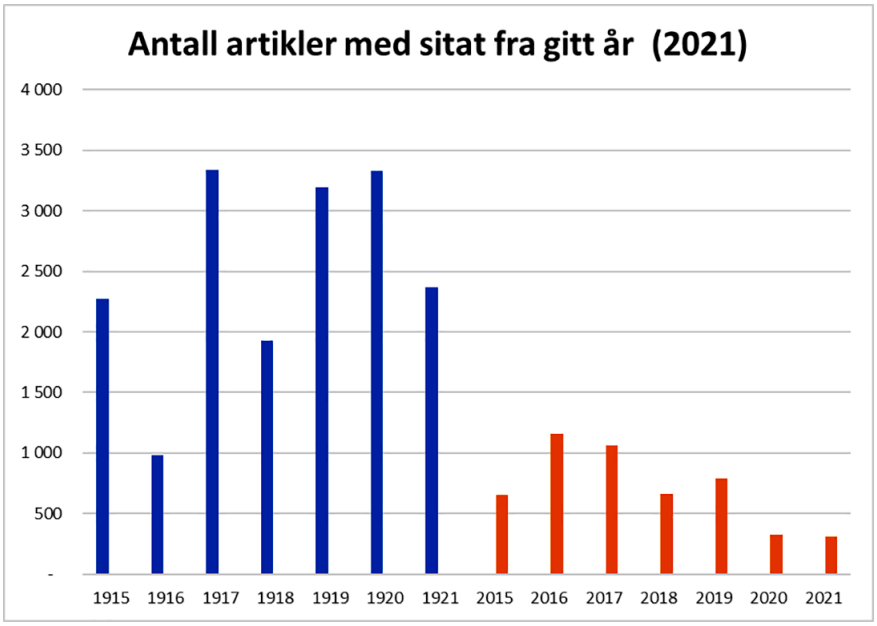
1. Innledning

Det Norske Akademis ordbok (NAOB; https://naob.no) er en historisk ordbok som beskriver bokmålet og riksmålet fra tidlig 1800-tall og frem til i dag. NAOB er også en litterær ordbok, slik at artiklene er bygget opp med utgangspunkt i den bruk som finnes i skriftlige kilder. Dessuten er NAOB en samtidsordbok – artiklenes hodeopplysninger følger dagens språk (med tanke på staving, bøyning og uttale), og de redaksjonelle eksemplene tilhører alltid nåtiden. Skal NAOB fungere som samtidsordbok, må vi sørge for kontinuerlig påfyll av nytt ordstoff på alle nivåer: ord, betydninger og faste uttrykk. Videre må sitater fra nåtiden tas inn uten særlig forsinkelse og i et omfang som monner i en base med over 200 000 oppslagsord og mer enn 300 000 sitater.

NAOB finansieres dels av staten ved Kultur- og likestillingsdepartementet, dels av private stiftelser. For 2022 og 2023 har prosjektet fått til sammen 800 000 kroner fra stiftelsen Fritt Ord (https://frittord.no/) for å øke andelen sakprosa i NAOB og å øke andelen litteratur fra de aller siste årene. Redaksjonen består av tre leksikografer, en mindre del av den ene stillingen går til ledelse og administrasjon.

2. Endret siktemål

Når vi sier at NAOB beskriver norsk bokmål fra 1800-tallet og frem til i dag, er i dag bokstavelig ment: høyre ende av tidsaksen er åpen. NAOB er dokumenterende og litterær; vi dokumenterer ordforrådet gjennom den faktiske bruken i skrift. Ordbokens kildegrunnlag er åpent og utvides kontinuerlig. Samtiden skal være til stede på alle ordbokens nivåer: i lemmautvalget, i betydningsstrukturen og i kildene det siteres fra. Vårt forelegg er *Norsk Riksmålsordbok*, som har dokumentert store forfattere som blant andre de to Henrikene Wergeland og Ibsen svært grundig. Så var det også primært en resepsjonsordbok *Norsk Riksmålsordbok* var. NAOB er også en resepsjonsordbok, men vi har i større grad enn våre forgjengere et mål om at NAOB skal speile *språket*, og at samtiden hele tiden skal være present i ordboken. Den enkelte forfatter spiller en mindre rolle i NAOB, mens man i *Norsk Riksmålsordbok* dokumenterte enkelte forfatterskap nærmest uttømmende. Nå er det entydig språket som er objektet som dokumenteres, og det gjøres ved hjelp av litterære sitater. Ingen moderne forfatter er så grundig representert som tilfelle er med Wergeland og Ibsen. Med en så grundig dokumentasjon av den eldre litteraturen kreves det en målrettet innsats for å synliggjøre samtiden.



FIGUR 1. Antall artikler i NAOB med sitater fra årene 1915–1921 og 2015–2021. Tallene er hentet i 2021.

Figur 1 viser hvordan antall artikler med sitater fra årene 1915–1921 er langt høyere enn tallene hundre år senere. Tallene er fra før det nye prosjektet med å øke antall sitater fra de siste årene ble påbegynt. Det er ikke noe mål at alle år skal være likt representert, og heller ikke at årstallene skal «fylles opp» med det samme, for vi legger stadig til sitater også fra eldre litteratur. Men antall sitater fra senere år bør noe raskere kunne bli en del flere.

Ved inngangen til 2022 hadde ordboken 5000 sitater fra tidsrommet 2015–2021. Ved årets slutt var det 7700 sitater fra tidsrommet 2015–2022. Høyst er antallet for 2021, hvor det per mars 2023 er over 2000 sitater. Endringene kan leseren selv finne frem til ved å fritekstsøke årstallet man vil undersøke på naob.no.

2.1. Ulik bruk

Hvis en bruker søker opp ord av typen *stjerneblomme* og *begerbluss* vil hen naturlig nok møte sitater fra fortiden, og bare det. Søkes det på ord ord som *snitche* og *tæsje*, er det moderne kilder som er sitert. Sånn må

det være. Men søker man opp ord fra kjerneordforrådet, skal aktualiteten fremgå gjennom litteraturen som er sitert. Som samtidsfokusert NAOB-bruker skal man gjennom bruk alene kunne vite at NAOBs korpus ikke er lukket, ved at man stadig treffer på nyere årstall i artiklenes sitatseksjon. Vi har ingen ambisjoner om å finne det eldste belegget for hvert ord i NAOB, men vi skal hele tiden ha ordbokens tidsspenn med oss i redigeringen. Ordet *fullvaksinere* kom inn i NAOB i 2021, men det viste seg at det ikke var nytt. Et eldre sitat kom derfor også med. Som oftest er det motsatt: Vi har artikler som har vært med helt fra *Norsk Riksmålsordbok*, med eldre sitater. Da er det viktig at nyeste sitat for et fullt brukbart ord ikke er fra for eksempel 1960-tallet. Det er utenkelig å komme igjennom hele basen i løpet av få år, men gjennom Fritt Ord-prosjektet vil vi få sitatfornyet artiklene for en stor mengde aktuelle ord.

3. Kildegrunnlaget og basene det aksesseres fra

I Norge ligger det meste av bøker utgitt frem til år 2000 digitalt og fritt tilgjengelig hos Nasjonalbiblioteket (www.nb.no), men hva så med de siste 22 årene? Vi kan ikke nøye oss med sitater fra aviser og tidsskrifter når vi skal vise språket i bruk de to siste tiårene. Det skapes lett et inntrykk av at NAOB er tilbakeskuende, med den store mengden sitater fra Ibsen og andre 1800-tallsforfattere vi har med oss fra *Norsk Riksmålsordbok*. NAOB-redaksjonen ønsker i de kommende årene å legge enda større vekt på å få inn sitateksempler fra ny litteratur tidlig. I det følgende vil jeg vise hvordan redaksjonen arbeider for å ha på plass et representativt utvalg av ny litteratur allerede i utgivelsesåret og vise hvordan vi har lagt opp dette arbeidet for å få mest mulig samtidsordbok for pengene.

NAOBs litteraturliste består av godt over 6000 bøker og rundt 800 aviser og tidsskrifter (https://naob.no/litteratur). Vi har ikke noe korpus som rommer dem alle, så hvordan når vi våre kilder?

For aviser og tidsskrifter bruker vi A-tekst – et abonnementsbasert mediearkiv med over 300 aviser og tidsskrifter, eller vi kan også gå rett til kilden, til avisens nettutgave. For sakprosa og skjønnlitteratur har vi flere gode baser, men de fleste har et litteraturutvalg som stopper i år 2000, noen også før. Litteraturen fra de manglende 22 årene kan vi som dokumenterende ordbok ikke unnvære. Vi skal se nærmere på to av basene, og også se på hvordan vi får tak i den nyeste litteraturen.

3.1. Nettbiblioteket

Nettbiblioteket (tidligere kalt Bokhylla) er en digital tjeneste hos Nasjonalbiblioteket som tilbyr søk i norske bokutgivelser, fritt tilgjengelig. Nettbiblioteket er en språkskatt og en ressurs vi bruker mye. Fordelene er mange: Det enorme omfanget er svært nyttig siden vi også beskriver forholdsvis sjeldne ord, aviser og fagtidsskrifter samlet på ett sted effektiviserer arbeidet og nettstedet har mange funksjoner, eksempelvis N-gramsøk (https://www.nb.no/ngram/). Haken ved Nettbiblioteket er at det som er fritt tilgjengelig for allmennheten, er utgivelser som er fra år 2000 eller eldre, med unntak av noen verk som rettighetshavere har holdt tilbake. Stoff fra etter år 2001 er kun tilgjengelig med en særskilt tilgang som bare tildeles utdanningsinstitusjoner. NAOB, som ikke er tilknyttet en utdanningsinstitusjon, får det ikke. Men det finnes andre gode løsninger.

3.2. DocFetcher

DocFetcher er en "open source desktop search application", altså programvare som lar deg søke i en mengde av filer og returnerer filene som inneholder søkestrengen (http://docfetcher.sourceforge.net/en/index.html). I praksis blir dette et korpus. Søkemessig er det ganske likt som Nettbiblioteket: Man kan trunkere og gjøre flerordssøk, men man får ikke opp noen konkordans. Vi har altså liten nytte av å fylle DocFetcher med filer som allerede er fritt tilgjengelige, for dem når vi gjennom Nettbiblioteket. Det vi har behov for, er filene med utgivelser fra 2001 og senere. Vi har derfor kontaktet forlagene direkte, fortalt dem om vårt behov og bedt om å få tilsendt utvalgte PDF-filer. Disse filene legges inn i DocFetcher. Vi velger ut forfattere og bøker som preger samtiden, og vi velger ut sakprosa, både ut fra bøkenes utbredelse og ut fra fag/tema. Vi vektlegger også bevisst en balanse mellom mannlige og kvinnelige forfattere. Forlagene er velvillige og sender oss de filene vi ber om. Både forfattere og forlag bidrar med glede til vårt prosjekt; de ser viktigheten av dokumentasjons- og formidlingsarbeidet. Vi har fått på plass en standardkontrakt, slik at forlagene er trygge på at vi ikke videredistribuerer filene.

For vårt arbeid er det en ulempe at Nettbiblioteket ikke kan skille ut oversatt litteratur (som ikke er del av NAOBs tekstgrunnlag). Med det skreddersydde DocFetcher-korpuset er det problemet løst. Denne avgren-

singen gjør også at hver kilde blir mer brukt. DocFetcher har videre den fordel at vi kan navngi filene og utnytte dette i sorteringen av treffene. For bøker vi ønsker å sitere ekstra mye fra, kan filnavnene starte med 01. Disse filene legger seg da øverst når trefflisten sorteres på filnavn. DocFetcher-korpuset brukes i den generelle redigeringen, men også mer målrettet. Den generelle redigeringen vil si når vi arbeider med en ny artikkel, eller arbeider med å få inn nye sitater i eksisterende artikler. Da kan vi bruke kilder med prefiks 01 i navnet for å styrke andelen sitater fra nyere litteratur eller sakprosasitater. Målrettet redigering er det f.eks. når vi i fagspråkprosjektet skal styrke konkrete fag eller domener. Da kan vi sanke både ord og sitater fra DocFetcher-korpuset.

Med DocFetcher får vi altså arbeidet svært effektivt med begge Fritt Ord-prosjektene samtidig; ny litteratur finner veien inn i NAOB, både i form av sakprosa og skjønnlitteratur. Ordforrådet kan finnes i NAOB fra før og bare suppleres med sitat, nytt ordforråd, nye betydninger eller nye faste uttrykk kan legges til.

4. Metodikk for sakprosaprosjektet

4.1. Ekserpering

Ekserpering er en svært nyttig metode for den type leksikografisk arbeid som utføres i NAOB, men det er en tidkrevende arbeidsmåte. Vi har såpass begrensede ressurser at vi ikke kan sitte og lese hele bøker i arbeidstiden. En mulighet er å streke under godt ordstoff i bøker man allikevel leser på fritiden, og så bearbeide fangsten i arbeidstiden. Dette er et svært nyttig tilskudd til ordboken, men det er på ingen måte forventet av leksikografene, og dermed ikke en metodikk vi kan basere ordbokens utvikling på. Utvalget blir dessuten tilfeldig, og dette er altså bare et tilskudd til det mer planmessige arbeidet.

Boka *Dyrespor* av Arnodd Håpnes fra 2021 tar for seg dyrespor og sportegn for alle ville dyr som finnes i Norge. Den favner felter som zoologi, biologi og økologi, og den er skrevet for allmennheten. Den type bøker er gode kilder for oss ved at de gir gode, informative sitater og viktig ordforråd. Det samme gjelder eksempelvis også pubertetsbøkene *Jenteboka* (Brochmann & Dahl 2019) og *Gutteboka* (Brochmann & Dahl 2021), som særlig dekker medisin og anatomi. I tillegg er det naturligvis alltid også mye allmennspråklig stoff i sakprosa som også siteres. Mye

sakprosa er det enklere å høste fra enn skjønnlitteratur, siden sakprosautgivelser ofte har en innholdsfortegnelse og kanskje også en ordliste. Dermed har man tilgang til sentrale ord fra faget eller domenet. Mange sitater kan også gjenbrukes. Et sitat som «fjellreven … beveger seg [som regel] i lett galopp, mens rødreven oftest traver» (Håpnes 2021:82) er like godt som belegg i alle de fire artiklene *rødrev*, *galopp*, *trave* og *fjellrev*. Tilsvarende har vi nylig tatt inn i litteraturlisten en trilogi om funksjonshemning, skrevet av Jan Grue. Her er det ingen innholdsfortegnelse eller ordliste, men vi kan søke opp domenerelevant stoff, for eksempel sammensetninger med *rullestol*. På den måten fikk *rullestolrampe*, *rullestoltilgjengelig* og *rullestoltilgjengelighet* en plass i NAOB. Dette igjen inngår i et annet av 2022s prosjekter, nemlig mangfold i ordboken. Prosjektet er tematisert i Hanne Lauvstads artikkel i denne konferanserapporten (Lauvstad 2023). Med PDF-filene har vi hatt god nytte av å foreta ekserpering ved målrettede søk ut fra bokens tema, ved gjenbruk av sitater og gjennom skumlesing, og dessuten ved bruk av språkteknologisk hjelpemidler.

4.2. Verktøyet Sketch Engine

Det mer manuelle arbeidet har altså fortsatt mye for seg i NAOB-redigeringen, men språkteknologiske hjelpemidler er selvsagt også viktige. Med PDF-filene for hånden, er Sketch Engine et ypperlig verktøy (https://www.sketchengine.eu/). Der kompilerer man egne korpus på nærmest ingen tid, fra filer man selv besitter. I sakprosa-prosjektet er det nyttig å lage korpus av én og én bok, eller av flere bøker som tilhører samme fag eller område. Disse sammenlignes så mot et mye større korpus for å skille ut tekstens nøkkelord, det vil si de ordene som preger teksten ved å ha høyere frekvens enn i sammenligningskorpuset. Det man ikke får høstet rikt av på denne måten, er allmennspråket i sakprosaen, fagrelevant stoff uten høy frekvens eller syntaktiske konstruksjoner, som burde vises frem i ordboken. Der kan skumlesing gi en viss fangst, og også Sketch Engines kollokasjons-funksjon. Sitatet «ulvemotstanderne har en til dels sterkt utviklet motvilje mot forskere, som ofte oppfattes som naturvernernes, forvaltningens og fagstatsrådens forlengede arm» (Rossavik 2021:69), som på en fin måte viser uttrykket *noens forlengede arm*, vil ikke tre frem i en word sketch-analyse, mens leksikografblikket straks vil gjenkjenne sitatets illustrative kvaliteter.

4.3. Utvelgelse av litteratur

Redaksjonen arbeider i fellesskap med å plukke ut ny litteratur. For skjønnlitteratur vektlegger vi bøker som har fått en stor leserkrets og/eller viktige priser, og således har preget tiden. Eksempler er Linn Ullmanns *Jente, 1983* (Ullmann 2021), Trude Marsteins *Egne barn* (Marstein 2022), Jan Kjærstads *En tid for å leve* (Kjærstad 2021). På den annen side gjør vi også en innsats for å løfte frem mindre etablerte navn og sjangere som får mindre oppmerksomhet. Barne- og ungdomslitteratur, både fra samtiden og eldre, ønsker vi generelt å dekke bedre. Sakprosa velges både ut fra fagområder hvor NAOB trenger styrking, hvor det har skjedd mye nytt og ut fra hva som preger allmennheten og dermed allmennspråket.

5. Avslutning

NAOB skal være oppdatert når det gjelder både ordforrådet og litteraturen, og dette henger sammen. I sakprosa-prosjektet skal vi styrke ordforrådet og mengden sitert litteratur generelt, i tillegg til at vi har sett oss ut noen områder som er ekstra aktuelle gjennom sin plass i allmennheten, blant annet økologi og medisin. Dette er fagområder som trenger mer og mer inn i allmennspråket, og hvor det er behov for gode leksikografiske beskrivelser, slik man kan forvente å finne det i en samtidsordbok. Samtidig vokser bekymringen for norsk som fagspråk fordi engelsk blir stadig mer dominerende. Fagmiljøene selv må bruke norsk og gjennom bruken opprettholde og utvikle norsk fagspråk. En dokumentasjonsordbok som NAOB er en viktig støttespiller i dette arbeidet. Fagspråk er pekt ut som satsningsområde for ordboken i de nærmeste årene, og her er samtiden en forutsetning.

Litteratur

Ordbøker

NAOB = *Det Norske Akademis ordbok*. Oslo: Det Norske Akademi for Språk og Litteratur. <http://naob.no> (mars 2023).

Annen litteratur

Lauvstad, Hanne (2023): Sensitive ord i *Det Norske Akademis ordbok*. Utfordringer i diakron leksikografi. I: Holmer, Louise et al. (red.), *Nordiska studier i lexikografi 16*. Lund & Göteborg: Nordiska föreningen för lexikografi, 213–224.

Sitert og omtalt kildelitteratur

Brochmann, Nina og Dahl, Ellen Støkken (2021): *Gutteboka. Din guide til puberteten*. Oslo: Aschehoug.

Brochmann, Nina og Dahl, Ellen Støkken (2019): *Jenteboka. Ellen og Ninas guide til puberteten*. Oslo: Aschehoug.

Grue, Jan (2018): *Jeg lever et liv som ligner deres: En levnetsbeskrivelse*. Oslo: Gyldendal.

Grue, Jan (2021): *Hvis jeg faller: En beretning om usynlig arbeid*. Oslo: Gyldendal.

Grue, Jan (2022): *Prøve og feile*. Oslo: Gyldendal.

Håpnes, Arnodd (2021): *Dyrespor. Og andre tegn etter dyr*. Oslo: J.M. Stenersens forlag A.S.

Kjærstad, Jan (2021): *En tid for å leve*. Oslo: Aschehoug.

Marstein, Trude (2022): *Egne barn*. Oslo: Gyldendal.

Rossavik, Frank (2021): *Ulv? Ulv! En bok om rovdyr og mennesker i Norge*. Oslo: Cappelen Damm.

Ullmann, Linn (2021): *Jente, 1983*. Oslo: Forlaget Oktober.

Et fullformsystem for analyse av eldre tekst på tidlig nynorsk, bygd på Aasen-normalen

Christian-Emil Smith Ore, Oddrun Grønvik & Trond Minde

Many European languages have undergone considerable changes in orthography over the last 150 years. This hampers the application of modern computer-based analysers to older text, and hence computer-based annotation and studies of text collections spanning a long period. In 2021, as a step towards a functional analyser for Norwegian texts (Nynorsk standard) from the 19th century, we did a pilot project with the objectives to create a full form generator for all inflected forms of headwords found in Ivar Aasen's dictionary published in 1873 (Aasen 1873) and based on the description in his grammar from 1864 (Aasen 1864).

As a test, the full form list generated from this new word bank was used to analyse the word inventory of texts by A.O. Vinje, written in the period 1850–1870. The Vinje texts were also analysed using a full form list of modern standard Norwegian, to study the differences in applicability and see how Vinje's language relates to the written standard of modern Norwegian.

KEYWORDS: digital lexicography, full form systems, 19th century, orthographic history, text analysis

1. Bakgrunn

Norsk Ordbok har som andre store nasjonalordbøker vært et langvarig prosjekt. Det ble påbegynt rundt 1935 og redigeringen ble avsluttet i 2015. I 2002 ble det norske kulturdepartementet og Universitetet i Oslo enige om en restrukturering av prosjektet med nødvendig finansiering under forutsetning av at ordboka ble ferdigstilt til grunnlovsjubileet i 2014. Det holdt nesten, og ordboka forelå i 12 bind i mars 2016. I dette prosjektet, også kalt NO2014, redigerte en alfabetdel I–Å og halve H. Det er i de senere årene satt i gang et prosjekt for å revidere A–H. En av forutsetningene for NO2014-prosjektet var at ordboka skulle redigeres digitalt på en slik måte at den også kunne publiseres som en digital ordbok sammen med det viktigste grunnlagsmaterialet. Store deler av dette grunnlagsmaterialet, blant annet seddelsamlingen på 3,2 millioner sedler, var allerede

digitalisert på 1990-tallet (Ore 1998). Det digitale materialet ble koplet til redigeringsdatabasen via Metaordboka (Ore 1998, 2012; Ore & Grønvik 2018), som er et verktøy for å systematisere leksikografisk kildemateriale og er spesielt egnet for dokumentasjon av svakt normerte språk som dialekter og språk med stor ortografisk variasjon. I tillegg til Norsk Ordboks seddelsamlinger og nyordsamlingen for bokmål inneholder Metaordboka et digitalt ordbokbibliotek, med 95 store og små ordbøker.

Metaordboka er et register over det norske leksikonet der hver artikkel representerer en leksikalsk gjenstand (Atkins & Rundell 2008), utvidet til å omfatte ortografiske varianter. I tillegg til bruksbeleggene har en metaordboksartikkel en oppslagsdel med redaksjonelle varianter. For hver variant er språk (bokmål/nynorsk), ordklasse, rettskrivningsstatus og tidsrom for denne statusen angitt. I tillegg er sammensetninger leddanalysert. Det er for tiden om lag 780 000 artikler i Metaordboka. Vi har i de siste seks årene sett på hvordan den kan brukes til dokumentasjon og analyse av skriftspråkutviklingen i Norge de siste 150 årene. Et første mål i dette arbeidet var å se i hvilken grad vi kunne bruke kildene i Metaordboka til å samordne ordtilfanget i skriftmålsstandardene bokmål og nynorsk (Ore & Grønvik 2018a), og deretter å se om ordbokbiblioteket i Metaordboka viste utviklingen av et kjernevokabular for bokmål (Ore & Grønvik 2018b). Siden en rekke offisielt godkjente ordlister og ordbøker er knyttet til Metaordboka, danner den et skjelett for en historisk normordbok. *Det Centrale Ordregister*, som ble presentert på NFL 2022 (Henrichsen 2023), har mye av den samme funksjonaliteten som Metaordboka, men har et litt annet fokus, struktur og implementasjon. Men det kan være interessant å se på de to i sammenheng siden norsk bokmål og dansk har en felles historie til rundt 1900.

Alle ressursene og infrastrukturen som er nevnt i denne artikkelen, vedlikeholdes nå av *Språksamlingane* ved Universitetet i Bergen. En del av resultatene i denne artikkelen er også beskrevet i Ore, Grønvik & Minde (2022).

2. Norsk ordbank

I tillegg til Metaordboka kommer Norsk ordbank, som er en oversikt over grunnord og paradigmer samt fullformgeneratorer for bokmål og for nynorsk. Den bygger på IBMs stavekontroll fra 1980-tallet og arbeid

gjort i forbindelse med et prosjekt for å lage en morfo-syntaktisk, regelbasert tonivåtagger (Koskenniemi 1990) for moderne norsk (Hagen et al. 2000). Grunnordene i ordbanken er lenket til oppslagsordene i Bokmålsordboka (Wangensteen 1986–2006, Ordbøkene 2023) og Nynorskordboka (Hovdenak et al. 1986–2006, Ordbøkene 2023). Oppslagsord og bøyingsinformasjon i ordbøkene hentes direkte fra ordbanken (se https://ordbøkene.no), og Språkrådets anbefalte rettskrivning formidles via ordbanken.

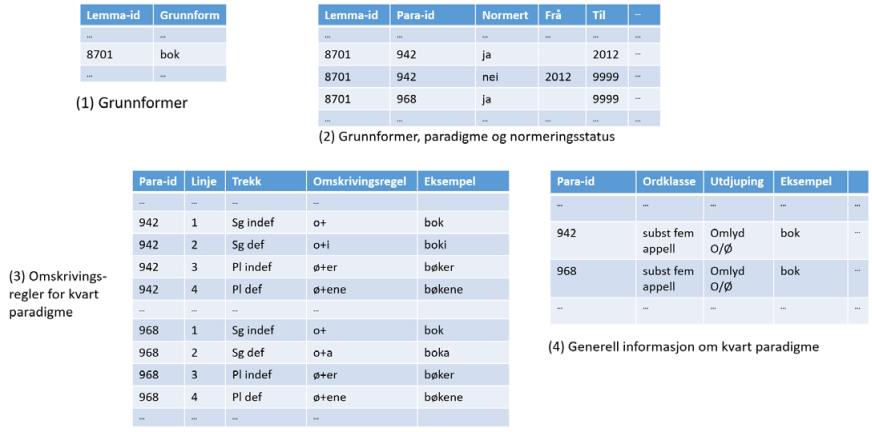
I vårt arbeid forstår vi med en ordbank en datastruktur for å lagre informasjon om ord (leksikalske gjenstander) og deres bøyde former. Den grunnleggende ideen er at et ord (leksikalsk gjenstand) er identifisert som mengden av alle dets mulige bøyde former. I denne modellen vil oppslagsformen i en ordbok kunne brukes som en representant for hele mengden av bøyde former. Både for bokmål og nynorsk har det vært mange ganske gjennomgripende revisjoner siden slutten av attenhundretallet.

Figur 1 viser en forenklet versjon av en ordbank. Strukturen består av (1) en liste av grunnformer (oppslagsord), (2) en tabell med informasjon om hvilket eller hvilke paradigmer grunnformen kan følge, og (3) en liste av omskrivningsregler (paradigmer) for å generere de bøyde formene fra grunnformen. Tabellen som vises i (2), er krumtappen i systemet. For en gitt grunnform vil en linje i tabellen fortelle hvilket eller hvilke paradigmer grunnformen følger, den ortografiske status og tidsrommet for denne. I nynorsk hadde for eksempel ordet 'bok' inntil 2012 sideformen 'boki' i tillegg til 'boka' i bestemt form entall. Siden 2012 har formen 'boka' vært eneste tillatte form. Paradigmetabellen (3) i figur 1 viser noen av de omskrivningsreglene som brukes for å lage fullformene. Omskrivningsprosessen fungerer som følgende: En grunnform sammenliknes med omskrivningsreglen i linje 1. I eksempelet i figur 1 vil jokertegnet '+' bli bundet til 'k'. Det vil si at det konstante initialsegmentet vil være 'b', mens finalsegmentet som endres, er 'ok'. Linjene 2 til 4 angir endringene som må gjøres for å lage fullformlistene 'bok, boki, bøker, bøkene' og 'bok, boka, bøker, bøkene'. Det kan være stort overlapp mellom fullformlistene generert av ulike paradigmer for ett og samme ord slik eksempelkolonnen viser. I en morfologisk analysator vil en som oftest være interessert i å finne grunnformen (lemma) og hvilken paradigmelinje (morfologiske trekk) formen representerer. En vil derfor slå sammen slike overlapp og lage en liste uten duplikater av parene (fullform, morfologiske trekk). Sys-

temet med omskrivningsregler er basert på en stavekontroll utviklet av IBM på slutten av 1980-tallet (Engh 2014).

Hvert paradigme har i tillegg til det som er vist i figur 1, informasjon om ordklasse, prototypiske grunnformer og eventuelle kommentarer. I Aasen-ordbanken har en for hvert paradigme referanser til de relevante paragrafene i Aasens grammatikk (Aasen 1864).

En tilleggsgevinst er at ordbankene dokumenterer endringene i rettskrivningen foreløpig siden 1995. Siden all kopling mellom grunnord og paradigmer er angitt med normeringsstatus og tidsrom, kan en hente ut rettskrivningsnormen slik den var på et gitt tidspunkt.



FIGUR 1 En skisse av strukturen i ordbanken for moderne nynorsk.

Ordbanken og Metaordboka har ulik historie og formål, men begge dokumenterer det norske ordinventaret. Ordbankens fokus er moderne norsk slik vi finner det i Bokmålsordboka og i Nynorskordboka. Metaordboka favner mye videre og dokumenterer ordinventaret tilbake til slutten av 1800-tallet. Ordbanken inneholder en fullformgenerator og genererer alle mulige bøyde former og markerer hvilke som er i samsvar med Språkrådets anbefalinger. En slik funksjon finnes ikke i Metaordboka.

3. Analyseverktøy for eldre tekst

Norsk ordbank er begrenset til moderne rettskrivning, og en analysator basert på ordbanken fungerer ikke særlig godt for tekster fra før de store

rettskrivningsreformene på 1930-tallet. Om en kunne utvide fullformgeneratorene til å omfatte nå foreldede bøyingsformer av grunnord i eldre rettskrivning, vil en kunne bygge analysatorer med bedre dekningsgrad. Det beste ville være å lage fullformer for alle oppslagsord i Metaordboka, men det reiser prinsipielle spørsmål. Om en skal lage en analysator for eldre tekst, må eldre ordformer kunne bøyes etter en antatt offisiell standard på tidspunktet for belegget, og hvordan skal en (re)konstruere eldre standarder, som vil ha hypotetiske former?

En vanlig strategi for å konstruere en hypotetisk standard er å gjennomføre en analyse av utvalgte tekster fra den perioden en vil arbeide med. På grunnlag av denne lager en så en oversikt over leksikalske gjenstander, hvordan de er realisert, og etablerer en tentativ norm med grunnformer og bøyingsformer. Dette er en standard metode for en leksikografisk beskrivelse av språk. En annen vei, som forutsetter at språket til en viss grad er dokumentert, er å bruke relevante, autoriserte ordlister og ordbøker som ligger nærmest opp til (helst før) den perioden en vil studere, og å bruke grammatikker fra samme periode til å sette opp bøyingsskjema. Denne metoden er godt egnet til å se i hvilken grad en (konstruert) norm er brukt i tekster fra den gitte perioden. Men et problem er at slike ordlister ofte har relativt få oppslagsord og dermed har en lav dekningsgrad. Dette er typisk for germanske språk med produktiv ordsammensetning. Den beste løsningen er å kombinere de to tilnærmingene, altså både analysere et korpus av eldre tekster og ta utgangspunkt i tilgjengelige grammatikker og ordlister. Begge krever mye arbeid, men kan gi et nyttig analyseredskap.

Som et første ledd i arbeidet med analysatorer for eldre bokmål og nynorsk har vi gjennomført et fullskala pilotprosjekt der vi tok utgangspunkt i det som danner begynnelsen for nynorsk (landsmål) som skriftspråkstandard, nemlig Ivar Aasens *Norsk Grammatik* (Aasen 1864) og *Norsk Ordbok med dansk Forklaring* (Aasen 1873). Til sammen danner de en beskrivelse av norsk folkemål og et forslag til en norsk rettskrivning.

Aasen var selvlært lingvist i en tid med mange selvlærte lingvister og en raskt voksende sammenlignende historisk språkvitenskap. Aasen studerte den danske lingvisten Rasmus Rasks arbeider og ulike språk som norrønt, latin, gresk og moderne språk. I 1840 fikk Aasen i oppdrag fra *Det Kongelige Norske Videnskabers Selskab* (DKNVS, Trondheim) å dokumentere det norske folkespråket gjennom dialektene og teste hypotesen om at norsk språk stammet fra norrønt og ikke var et forvansket dansk (Aasen

1848 og 1850, Walton 2016). Å dokumentere norsk folkespråk var en oppgave han fortsatte med resten av livet. En skal være klar over at han skrev for forskersamfunnet, og at han vurderte å skrive ordforklaringene i ordbøkene på tysk for å nå et større publikum, men endte med danske ordforklaringer. Formålet var altså å dokumentere norsk folkemål og ikke å lage en ordbok til støtte i daglig tekstproduksjon. Det hadde konsekvenser for ordutvalget i ordboka. Nyere importord ble stort sett utelatt. Vi har gitt en bredere beskrivelse av Ivar Aasen og hans ordbok og grammatikk (Aasen 1873 og Aasen 1864) i det andre bidraget på NFL 2022 (Grønvik, Ore & Minde 2023).

Grunnlagsmaterialet for pilotprosjektet er Aasens ordbok (1873) og grammatikk (Aasen 1864). Ordboka ble skrevet inn som formatert digital tekst i 1996. Den ble publisert i en CD-utgivelse og senere på Nynorsk kultursentrums nettsider, og var grunnlaget for Aarset og Krukens trykte versjon fra 2003 (Aasen 2003). I 2016 ble den digitale teksten (semi-)maskinelt analysert og gitt en TEI-koding, se Ore (2016). Vi ønsket bare å finne oppslagsord Aasen mente hørte til en mulig standard. I tillegg til uthevede ord på typisk oppslagsplass tok vi bare med uthevede ord med ordklassemarkering. Analysen resulterte i om lag 36 000 artikler og 42 000 oppslagsord. I tillegg kom 10 000 dialektformer som står i den løpende teksten. Den senere manuelle gjennomgangen, har resultert i 7 000 ekstra oppslagsord. Dette skyldes tekstlige og språklige forhold som det er gjort nærmere rede for i vårt andre bidrag (Grønvik, Ore & Minde 2023). Aasens grammatikk foreligger også som digital tekst og er tilgjengelig på nettsidene til Nynorsk kultursentrum, men det er ikke foretatt noen videre oppmerking eller analyse av teksten.

Vi bestemte oss for å lage en egen ordbank for Aasen-standarden og holde den separat fra ordbanken for nynorsk. På et senere tidspunkt kan en eventuelt lage en felles ordbank over nynorsk fra 1864 til i dag. Den største forskjellen mellom ordbanken for moderne nynorsk og den for Aasen-normalen finner en i paradigmene. Aasen-standarden opererer med flere morfologiske trekk, og paradigmene har dermed flere linjer, slik figur 2 viser.

P-id	Linje	Omskrivings-regel	Trekk	forklaring	Fullform
942	1	o+	eint ub	Eintal ubunden	bok
942	2	o+i	eint bu	Eintal bunden	boki
942	3	ø+er	fl ub	Fleirtal ubunden	bøker
942	4	ø+erna	fl bu	Fleirtal bunden	bøkerna
942	5	o+enne	dat eint	Dativ eintal	bokenne
942	6	o+om	dat fl	Dativ fleirtal	bokom
942	7	o+ar	gen eint	Genitiv eintal	bokar
942	8	o+a	gen fl	genitiv fl	boka

P-id	Ordklasse	Utdjuping	Eksempel	Grunngjeving
…	…	…	…	
942	subst fem appell	OmlydV O/Ø	bok	Aasen 1864 §171
…	…	…	…	

FIGUR 2 Et paradigme for feminine substantiv med O/Ø-omlyd

For hvert paradigme i Aasen-ordbanken er det en referanse til den eller de paragrafene i Aasens grammatikk som begrunner paradigmet. Aasen-ordbanken har i alt 412 paradigmer mot den moderne ordbankens 580. Dette skyldes nok at den moderne ordbanken har en rekke spesialparadigmer for enkeltord. Ser vi på antall grunnord, har Aasen-ordbanken om lag 46 000 og den moderne 105 000. Av genererte entydige fullformer er det rundt 480 000 i Aasen-ordbanken og 580 000 i den moderne. Så vi ser at den rikere Aasen-morfologien genererer relativt sett flere ordformer.

4. En første dekningstest

Som en første test på dekningsgraden ønsket vi å kjøre fullformlistene fra Aasen-ordbanken og den moderne ordbanken mot et sett av tidlige nynorsktekster og valgte en samleutgave av verkene til den norske forfatteren og journalisten A.O. Vinje (1818–1870). Vinje døde tre år før Aasens ordbok (Aasen 1873) kom ut, så han kan ikke ha brukt Aasen 1873. Men de hadde i mange år tett kontakt. En antar at Aasen ga Vinje råd om rettskrivningen han skulle bruke, men han hadde også sine egne meninger om det. De viktigste valgene Vinje gjorde, ble publisert som et lite tillegg i hans tidskriftet *Dølen* (Vinje 1859). Vinje skrev for det meste artikler og essay. På midten av 1800-tallet var det i Norge ennå ikke noen tradisjon for å skrive lengre litterære tekster som romaner, noe som skiller Norge fra land som Frankrike og Storbritannia. Vinjes egne tekster i de samlede verkene utgjør ikke mer enn 590 000 løpende ord når de danske tekstene og kommentarapparatet holdes utenfor.

TABELL 1. Resultatet av å kjøre ordbanklistene mot Vinjes skrifter. Det er mange bøyingshomografer og ingen disambiguering, så det som ikke er funnet, er det mest interessante.

	Alle ordformer (tokens)		**Entydige ordformer (types)**	
I begge:	460 674	78 %	10 324	25 %
bare hos Aasen:	32 110	5 %	4 254	10 %
bare nynorsk (2012):	31 261	5 %	6 114	15 %
I alt:	524 045	89 %	20 692	51 %
Ikke funnet:	65 561	11 %	20 069	49 %
Ordformer i tekstene i alt	589 606		40 761	

Testen vi utførte, ble satt opp som følger: Fra Aasen-ordbanken brukte vi en liste med 520 000 entydige tripler (ordform + POS + informasjon om morfologiske trekk) som bestod av 290 000 entydige ordformer. En tilsvarende liste med 585 000 tripler ble lagd fra ordbanken for moderne nynorsk. Den utgjorde 416 000 entydige former. Her var utvalget begrenset til de anbefalte formene i gjeldende rettskrivning. Et lite, håndlagd konkordansprogram gikk gjennom teksten, isolerte ordformer og sjekket dem mot listene. Resultatet er vist i tabell 1.

Det totale antallet av entydige ordformer (types) i Vinje-tekstene er om lag 40 000. Av disse forekommer 23 360 en gang, mens 146 forekommer mer enn 500 ganger, noe som er i rimelig samsvar med Zipfs lov (Wikipedia 2023). Av de 146 ordformene med høyest frekvens finnes 128 i begge ordbankene, 8 bare i Aasen-ordbanken, 3 bare i nynorskordbanken, og 7 er Vinje-spesifikke. Av dem som bare forekommer i Aasen-ordbanken, er seks bøyde former som har endret staving etter 1873, og de to siste er grunnformer. De tre som bare forekommer i nynorskordbanken, er bøyde former som er blitt endret etter 1873. Flere av de Vinje-spesifikke formene betraktes i dag som Vinjes signaturformer. Han brukte 'ikki' istedenfor Aasens 'ikkje', og preposisjonen 'af' brukte han både i norske og danske tekster. Forekomstene av 'ikke' skyldes dels at det står noe dansk igjen i teksten vi har analysert, men også vakling hos forfatteren. Andre avvikende former kan skyldes at han hadde andre ortografiske vaner enn Aasen.

Omtrent 51 % av typene er funnet i én av eller begge ordbankene, mens 49 % ikke finnes i noen av dem. Vi har gjort en mer detaljert klassifikasjon av ordformene på 'a-' (i alt 1 449). Nedenfor har vi kort kommentert disse ordformene:

1) Ordformene som er funnet i begge ordbankene, hører til kjernevokabularet i nynorsk (inkludert bøyde former). Antallet sammensetninger er relativt lite. Vi finner få importord siden Aasen oftest ikke tok disse med i ordboka. Det er få tvilstilfeller siden Aasens rettskrivningsnorm er veldig klart definert.
2) De ordformene som bare finnes i Aasen-ordbanken, er for det meste bøyde former som ikke lenger brukes i moderne nynorsk, for eksempel regelrette adjektiv som ender på «-ad». Endelsen er nå redusert til «-a».
3) Av de ordformene som bare finnes i Nynorsk-ordbanken, er 40 % ikke-germanske importord, 17 % navneformer og om lag 10 % danske ordformer som i dag er blitt en del av nynorsk. De resterende 33 % er former av grunnord som er i samsvar med Aasens rettskrivning, men mangler i ordboka (28 %) eller hører til Vinjes egen og mer heterogene rettskrivning (5 %). Vinje publiserte tekster to ganger i uken, og når han var i tvil, ser det ut til at han valgte ordformer fra hans egen dialekt i Vest-Telemark. Noen av disse har senere blitt tatt med i moderne nynorsk.
4) Ideelt sett burde flere av ordformene som ikke finnes i noen av ordbankene, vært analysert. Men ut fra analysen av ordene på «a-» kan vi med stor sikkerhet si at disse «Vinje-ordene» faller i følgende grupper: a) danske ordformer (fra sitater og kortere danske tekster), b) importord der mange nå har fått en fornorsket stavemåte, c) ordformer som er i samsvar med Aasens rettskrivning, men som ikke finnes i Aasens ordbok, d) ord som er særegne for Vinje, for eksempel ord som er ortofont stavet, og som ofte er påvirket av Vinjes egen dialekt, e) ord fra latin og moderne språk, særlig engelsk.

Vi kan konkludere med at Vinje og Aasen var ganske enige om hvordan en norsk rettskrivning skulle være. Blant de frekvente ordformene er det få uoverensstemmelser, men de er iøynefallende i Vinjes tekster siden det nettopp gjelder former som forekommer så ofte. Vinje var en forfatter og

journalist, og det viktigste for ham var å skrive slik at han ble lest. Om nødvendig lagde han nye ord, og i den prosessen virker det som norsk talemål var hans språklige kompass.

5. Videre arbeid

Norsk er i likhet med de andre skandinaviske språkene et germansk språk og har et produktivt system for ordsammensetning. Dermed er antallet entydige ord (typer) ubegrenset for alle praktiske formål. En fullformliste kan aldri være uttømmende. En løsning på dette problemet er å bruke en såkalt sammensetningsanalysator, altså et program som markerer skillet mellom mulige sammensetningsledd. For Aasen-prosjektet vil en sammensetningsanalysator være et neste skritt. Vi planlegger å teste en slik analysator som er utviklet for Oslo–Bergen-taggeren, se Hagen Johannessen & Nøklestad (2000) og Nøklestad (2022a, 2022b).

Videre er det nødvendig å utvide grunnordslisten for eldre nynorsk. På lengre sikt bør en også inkludere nyere rettskrivninger fra tidlig 1900-tall. For norsk bokmål fra 1800-tallet finnes det ikke noe klart startpunkt. En ordbank basert på ordforrådet i Molbech (1859) kunne være nyttig og kanskje bli et dansk-norsk samarbeidsprosjekt?

Referanser

Atkins, Beryl T. Sue & Michael Rundell 2008. *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.

Engh, Jan 2014. IBMs leksikografiske prosjekt for norsk 1984–1991. *Maal og Minne* 106:1, 67–101. <http://ojs.novus.no/index.php/MOM/article/view/225>. Hentet september 2022.

Grønvik, Oddrun 2016. The lexicography of Norwegian. I: Hanks, Patrick & Gilles-Maurice de Schryver (red.), *International Handbook of Modern Lexis and Lexicography*. Berlin, Heidelberg: Springer, 1–34.

Grønvik, Oddrun, Christian-Emil Smith Ore & Trond Minde 2023. Eit ikon kategorisert gjennom ein fullformgenerator. Om Ivar Aasens Norsk Ordbog med dansk Forklaring (1873). I: Holmer, Louise et al. (red.), *Nordiska studier i lexikografi 16*. Lund & Göteborg: Nordiska föreningen för lexikografi, 67–83.

Hagen, Kristin, Janne Bondi Johannessen & Anders Nøklestad 2000. A Constraint-Based Tagger for Norwegian. I: Lindberg, Carl-Erik & Steffen Nordahl Lund (red.), *17th Scandinavian Conference of Linguistics. Vol. I*. Odense: Institut for Sprog og Kommunikation, Syddansk Universitet, 31–47

Hagen, Kristin & Anders Nøklestad 2010. Bruk av et norsk leksikon til tagging og andre språkteknologiske formål. *LexicoNordica* 17, 55–72. <https://tidsskrift.dk/lexn/article/view/18624>. Hentet september 2022.

Henrichsen, Peter Juel 2023. Det Centrale Ordregister. Et indeks for det danske ordforråd – en gave til dansk sprogteknologi. I: Holmer, Louise et al. (red.), *Nordiska studier i lexikografi 16*. Lund & Göteborg: Nordiska föreningen för lexikografi, 113–126.

Hovdenak, Marit 2006 [1986]. *Nynorskordboka. Definisjons- og rettskrivingsordbok*. 4. utg. Oslo: Samlaget.

Koskenniemi, Kimmo 1990. Finite-State Parsing and Disambiguation. I: Karlgren, Hans (red.), *COLING'90. Proceedings of the 13th International Conference on Computational Linguistics. Vol. 2*. Stroudsburg: Association for Computational Linguistics, 229–232.

Molbech, Christian 1859. *Dansk Ordbog indeholdende det danske Sprogs Stammeord tilligemed afledede og sammensatte Ord, efter den nuværende Sprogbrug forklarede i deres forskiellige Betydninger, og ved Talemaader og Exempler oplyste*. København: Gyldendalske Boghandlings Forlag.

Nøklestad, Anders 2022a. *The Oslo-Bergen Tagger*. <https://github.com/noklesta/The-Oslo-Bergen-Tagger>. Hentet september 2022.

Nøklestad, Anders 2022b. *The Compound analyzer software*. <https://github.com/textlab/mtag>. Hentet september 2022.

Ordbøkene 2023. *Bokmålsordboka og Nynorskordboka*. <https://ordbokene.no/>. Hentet mars 2023.

Ore, Christian-Emil Smith 1998. Metaordboken – et rammeverk for Norsk Ordbok? I: Gellerstam, Martin, Kristinn Jóhannesson, Bo Ralph & Lena Rogström (red.), *Nordiska studier i lexikografi 5*. Göteborg: Nordiska föreningen för lexikografi, 250–270. <https://tidsskrift.dk/nsil/article/view/19466/17092>. Hentet september 2022.

Ore, Christian-Emil Smith 2012. Nettordbøker og Norsk Ordbok – hvordan etablere en vitenskapelig nettordbok. I: Eaker, Birgit, Lenn-

art Larsson & Anki Mattisson (red.), *Nordiska studier i lexikografi 11*. Oslo: Nordiska föreningen för lexikografi, 488–499. <https://tidsskrift.dk/nsil/article/view/19363/16988>. Hentet september 2022.

Ore, Christian-Emil Smith 2016. Gamle ordbøker og digitale utgaver. I: Gudiksen, Asgerd & Henrik Hovmark (red.), *Nordiske Studier i Leksikografi 13*. København: Nordisk forening for leksikografi, 203–216. <https://tidsskrift.dk/nsil/article/view/111227/160278>. Hentet september 2022.

Ore, Christian-Emil Smith & Oddrun Grønvik 2018a. Bokmål og nynorsk samindeksert – Metaordboka som verktøy for jamføring og utforsking av ordtilfang. I: Svavarsdóttir, Ásta, Halldóra Jónsdóttir, Helga Hilmisdóttir & Þórdís Úlfarsdóttir (red.), *Nordiske Studier i Leksikografi 14*. Reykjavík: Nordisk Forening for Leksikografi, 87–95. <https://tidsskrift.dk/nsil/article/view/117619>. Hentet september 2022.

Ore, Christian-Emil Smith & Oddrun Grønvik 2018b. Comparing Orthographies in Space and Time through Lexicographic Resources. I: Čibej, Jaka, Vojko Gorjanc, Iztok Kosem & Simon Krek (red.), *Proceedings of the XVIII EURALEX International Congress. Lexicography in Global Contexts*. Ljubljana: Znanstvena založba Filozofske fakultete Univerze v Ljubljani, 159–172. <https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/118/211/2923-1>. Hentet september 2022.

Ore, Christian-Emil Smith, Oddrun Grønvik & Trond Minde 2022. Word Banks, Dictionaries and Research Results by the Roadside. I: Klosa-Kückelhaus, Annette, Stefan Engelberg, Christine Möhrs & Petra Storjohann (red), *Proceedings of the XX EURALEX International Congress Dictionaries and Society*. Mannheim: IDS-Verlag, 321–333 <https://doi.org/10.14618/phpy-6r66>. Hentet mars 2023.

Ore, Christian-Emil Smith 2020. Å ta Hans Ross på ordet: Ross' ordbok i relasjon til Aasens med Metaordboka som verktøy. I: Sandström, Caroline, Ulla-Maija Forsberg, Charlotta af Hällström-Reijonen, Maria Lehtonen & Klaas Ruppel (red.), *Nordiska studier i lexikografi 15*. Helsingfors: Nordiska föreningen för lexikografi, 253–262. <https://tidsskrift.dk/nsil/article/view/124027/170987>. Hentet september 2022.

Vinje, Aasmund Olavsson 1916–1921. *Skrifter i Samling I–V*. Kristia-

nia: J.W. Cappelens forlag.
Walton, Stephen John 1996. *Ivar Aasens kropp*. Oslo: Samlaget.
Wangensteen, Boye 2005 [1986]: *Bokmålsordboka. Definisjons- og rettskrivningsordbok*. 3. utg. Oslo: Kunnskapsforlaget.
Wikipedia 2023. *Zipf's law* <https://en.wikipedia.org/wiki/Zipf%27s_law>. Hentet mars 2023.
Aasen, Ivar 1848. *Det norske Folkesprogs Grammatik*. Kristiania: Det Kongelige Norske Videnskabers-Selskab. Trykt hos Werner & Comp.
Aasen, Ivar 1850. *Ordbog over det norske Folkesprog*. Kristiania: Det Kongelige Norske Videnskabers-Selskab. Trykt hos Carl C. Werner & Comp.
Aasen, Ivar 1864. *Norsk Grammatik*. Christiania: Mallings Bogtrykkeri.
Aasen, Ivar 1873. *Norsk Ordbog med dansk Forklaring*. Christiania: Mallings Boghandel.
Aasen, Ivar 2003. *Norsk Ordbog med dansk Forklaring* (Ny utg. ved Kristoffer Kruken og Terje Aarset). Oslo: Samlaget

Kungliga Vetenskapsakademiens Handlingar som digitaliserad lexikalisk resurs. Fyra pilotstudier i historiskt akademiskt ordförråd

Lena Rogström & Sofie Johansson

In recent years, the amount of large, electronically published material has become still more interesting to lexicologists as well as lexicographers. However, it is still difficult, to perform systematic lexical studies of older texts since a morphosyntactic analysis is hampered due to the lack of a common orthographical norm.

In this article, four minor studies are described, all of which have been based on material from an important genre, which emerged during the 18th century, i.e. the scientific article. The articles have been published in the *Transactions of the Royal Academy of Sciences* and have been transliterated by the program *Transkribus*. In total, the material consists of 600.000 words.

The first study describes lexical establishment in the field of entomology; a scientific field first presented by Linnæus in 1739, which continued to develop in the discourse of *the Royal Academy of Sciences*. The second and third study compare the vocabulary of different fields of science with a modern academic vocabulary based on frequency analysis. The fourth study presents the lexical profiles of two prominent scientists of the 18th century, Linnæus and Charles de Geer. The results from these four minor studies indicate that both lexicology as well as historical lexicography would benefit from more research based on well structured corpora consisting of historical texts.

NYCKELORD: digitaliserad text, korpusanalys, historiskt material, akademiskt språk, lexikaliska verktyg

1. Inledning, syfte och disposition

I Sveriges största lexikografiska projekt, SAOB, vars första häfte kom ut 1893, har ordboksredaktionen under alla år samlat ihop, analyserat och publicerat lexikaliskt material på svenska från 1500-talet och framåt med omfattande uppgifter om ordens alla formella, etymologiska och semantiska sidor (se Lundbladh 1992; Nilsson 2019:25–29). Det gigantiska ordboksarbetet bygger på excerperingar av skriftligt material, vanligen i pappersform. Först på senare tid har redaktionen kunnat utnyttja elektro-

niska publikationer. Sedan 1997 finns själva ordboken tillgänglig digitalt, och den andra, reviderade upplagan kommer att bli en helt elektronisk ordbok. På så vis kan man säga att både den första och andra upplagan är typiska exempel på den utveckling inom såväl lexikografi som datateknik som präglat tidpunkterna för de båda upplagornas utarbetande.

De senaste åren har både tillgången till digitala textdatabaser samt förutsättningarna att digitalisera äldre textmaterial förbättrats, vilket resulterat i fler stora textbanker och elektroniska ordböcker. Textbanker och olika sökverktyg underlättar naturligtvis arbetet för lexikalisk forskning, men för historiska material är sökmöjligheterna fortfarande begränsade och inskränker sig ofta till pdf-ernas sökfunktioner. Material vars ortografi inte är normaliserad är extra svåra att söka i. Tillgången till elektroniska texter förenklar materialexcerperingen, men djupare lexikaliska analyser och jämförelser mellan olika tiders ordförråd är fortfarande svåra att göra. Språkbanken Texts korpusverktyg Korp (Borin et al. 2012b) erbjuder lemmatiserat material med ett flertal möjligheter till kombinatoriska sökningar, men för de historiska materialen är inte alltid alla funktioner tillgängliga.

Syftet med föreliggande artikel är att visa hur en digitaliserad korpus av historiska texter går att utnyttja både för kvalitativa och kvantitativa lexikaliska undersökningar. Materialet utgörs av artiklar från *Kungliga Vetenskapsakademiens Handlingar* (KVAH) (se avsnitt 3), och undersökningarna avser att ge exempel på hur man dels kan få en bättre överblick över ett specifikt ordförråd i sig, dels kan kartlägga lexikal förändring, vilket i sin tur går att utnyttja för både för lexikologer och lexikografer.

De frågeställningar som ligger till grund för pilotundersökningar är följande:

- Hur konventionaliseras beteckningar för nya eller förändrade begrepp inom enskilda fackområden?
- Hur ser framväxten av ett akademiskt ordförråd ut jämfört med moderna material?
- Kan man urskilja variation i det äldre akademiska ordförrådet inom olika ämnesområden?
- Vad säger enskilda författares lexikala profiler om framväxten av ett akademiskt ordförråd?

Avsnitt 2 utgörs av en kort bakgrund till valet av ämne och material. I avsnitt 3 beskrivs materialet och de generella metodiska utgångspunkterna. Undersökningarna beskrivs sedan i avsnitt 4 varvid mer specifika uppgifter om det metodiska genomförandet ges. Avsnitt 5 innehåller en summering av resultaten samt en framåtblick.

2. Bakgrund

Under 1700-talet utvecklas ett flertal nya genrer, bl.a. den naturvetenskapliga (Gunnarsson 2011b). Det akademiska språket och vetenskapliga genrer har idag stort inflytande på olika områden, och är intressanta att kartlägga. En viktig kanal för utvecklingen av moderna vetenskapliga genrer är den skriftserie som *Kungliga Vetenskapsakademien* (KVA) börjar publicera 1739 (Lindroth 1967:111–132). Akademien valde att publicera rönen på svenska, vilket varit viktigt för svenskt skriftspråk (Fries 1996:88). KVAH innehåller en stor mängd vetenskapsområden och skriftseriens storhetstid, 1739–1854, inramas av två giganter på naturvetenskapens område: Carl von Linné och Jöns Jacob Berzelius. KVAH har legat till grund för en del språkliga undersökningar (Gunnarsson 1987; Teleman 2011; Rogström 2019, 2023; Landqvist et al. 2020) men materialet erbjuder fortfarande en mängd infallsvinklar för undersökningar av den tidiga vetenskapliga prosan på svenska.

I våra studier görs åtskillnad mellan *vetenskapligt* och *akademiskt* ordförråd. Det vetenskapliga ordförrådet består av ämnesspecifika ord medan akademiska ord är sådana som används allmänt inom akademiskt språkbruk oavsett vetenskap. De är ofta svårare att använda korrekt eftersom de rent lexikaliskt kan ersättas med allmänspråkliga ord, varvid texten tappar sin speciella genretillhörighet (Sköldberg & Johansson Kokkinakis 2012:575).

3. Materialbeskrivning och generella metodiska utgångspunkter

KVAH finns idag digitaliserade i form av totalt 166 volymer som alla kan nås på internet. Texterna går att ladda hem både som bilder och i pdf-format, i vilka enkla ordsökningar kan göras.

Sedan några år tillbaka pågår ett arbete med att tillgängliggöra materialet i elektroniskt sökbart format. Under 2021 och 2022 genomfördes

ett mindre pilotprojekt med ekonomiskt stöd från Centrum för Digital Humaniora (CDH) vid Göteborgs universitet. Åtta volymer av KVAH digitaliserades i projektet (se tabell 1) av Johansson och Rogström. Urvalet representerar olika decennier, med en viss övervikt för 1740-talet (se avsnitt 4.3). Utöver de volymer som translittererats automatiskt tillkommer ett material som Rogström skrivit in manuellt och som uppgår till 83 rön om entomologi och 19 rön om åkerbruk. Också två handböcker om jordbruk och fåravel ingår i materialet. Sammanlagt uppgår hela materialet till 654 807 ord.

TABELL 1. Materialets sammansättning och storlek.

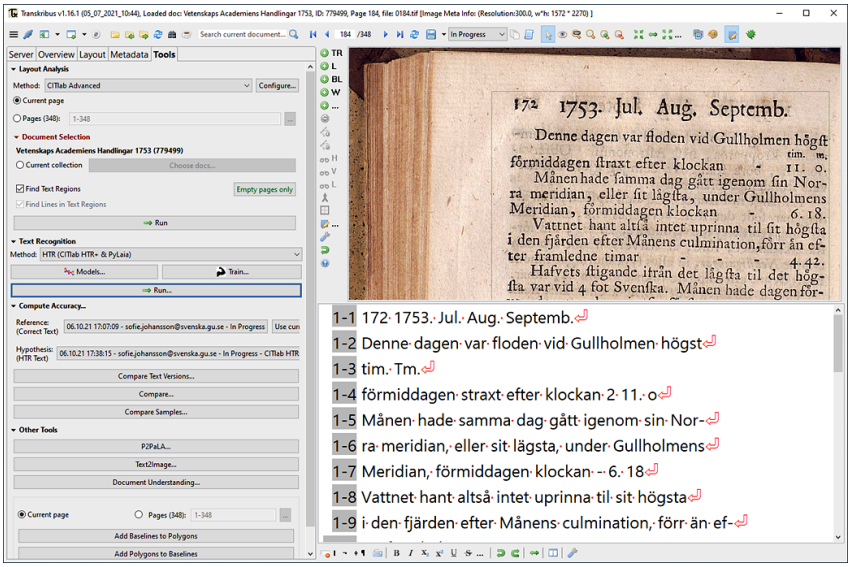
Text	Antal löpord
Åkerbrukstexter KVAH (1740–1748)	22 703
Stridsberg handbok (1727)	22 714
Serenius handbok (1727)	49 950
Entomologitexter KVAH (1739–1828)	116 707
KVAH, 8 volymer (1739–1774)	442 733
Totalt antal ord:	654 807

Det moderna jämförelsematerial som används utgörs av en akademisk ordlista (En svensk akademisk ordlista; jfr Sköldberg & Johansson Kokkinakis 2012) samt en balanserad referenskorpus, SUC 3.0. som finns fritt tillgänglig i Språkbanken Texts resursdatabas. Den akademiska ordlistan är baserad på akademiska texter vars ordförråd kategoriserats utifrån frekvens och spridning, varvid en frekventiell övervikt av ord i akademiska texter kunde urskiljas. Orden är gemensamma för flera ämnesområden och betraktas därför som ämnesneutrala akademiska ord (Johansson & Olander 2022). Materialet har främst använts för att analysera lärobokstext (se Johansson & Ohlsson 2019).

För omvandlingen av de digitala bilderna av KVAH till elektroniskt läsbar text har programmet *Transkribus* använts. Transkribus är ursprungligen ett verktyg för handskriven text, men har visat sig vara mycket effektivt också för tryckt text, inte minst frakturstil. Eftersom de fyra första volymerna i KVAH är tryckta i frakturstil var det lämpligt att transkribera alla dessa. Först genomfördes en inlärnings- och träningsfas baserad

på 20.000 ord. Därefter utvärderades hur väl programmet kände igen tecken och ord, och resultaten angavs i andel felprocent. Genom denna procedur skapades en språklig modell anpassad för det specifika textmaterialet. Felprocenten i analyserna i vårt material är 0,5 % på teckennivå och 2,07 % på ordnivå. Detta innebär att programmet identifierar enskilda tecken med större säkerhet än vissa ord. Alla rapporterade felprocent under 10 % anses vara mycket bra enligt utvecklarna. Alla texter har dessutom korrigerats manuellt.

Den lexikala analysen utfördes med programmen AntConc och AntWordProfiler i vilka det går att upprätta konkordanser och göra jämförelser av olika texters lexikala profiler.



FIGUR 1. Gränssnitt i *Transkribus* vid automatisk translitterering. Ur: Rön på Ebb och Flod vid Vårdhus och Nord-Caps-tracten, Af Anders Hellant (1753).

Textmaterialet saknar en gemensam ortografisk norm, vilket får konsekvenser när en dator ska analysera materialet. I analysen av mycket stora material får stavningsvariationen mindre betydelse, men för vårt material ger stavningsvariationen av vissa högfrekventa ord, såsom *och*/*ock*, utslag i statistiken. Den största variationen finns i volymerna fram till 1748. Mellan 1748 och 1773 trycktes KVAH av Lars Salvius som tillämpade en egen ortografisk norm (se Santesson 1986), och denna går att utnyttja i framtida bearbet-

ningar av materialet. I nuläget måste man dock ha i åtanke att stavningsvariation kan ha påverkat frekvensberäkningarna på materialet något, främst genom att ”samma” ord uppfattats som två olika om de stavats olika.

För jämförelsen med de akademiska ordlistorna har materialet lemmatiserats med hjälp av verktyget Sparv (Borin et al. 2016). Sparv möjliggör morfosyntaktisk analys av texter, vilket är användbart vid kartläggning och studier av språkbruk på ord-, fras- och meningsnivå. Sparv kan användas vid analys av äldre texter, men för text tillkommen före 1800-talet begränsar den ortografiska variationen användningsmöjligheterna. KVAH-materialet är i nuläget heller inte uppmärkt med någon metadata, såsom författare, årtal, genre och sidhänvisningar. Ett sådant arbete kommer att påbörjas under 2023 med finansiering av Svenska Akademien, varvid materialet blir ännu mer användbart. Materialet kommer i framtiden att publiceras i Språkbanken Texts korpusverktyg Korp som möjliggör korpusbaserade sökningar.

4. Fyra pilotundersökningar

4.1. Lexikala studier av ett specifikt fackområde – exemplet entomologi

I den första delstudien undersöks hur beteckningar för nya och delvis okända begrepp konventionaliseras. Många vetenskapsområden utvecklades snabbt, och KVA:s val av svenska som publiceringsspråk medförde ibland osäkerhet om vilka beteckningar som var lämpligast att välja, speciellt för nya upptäckter.

Ett område som började utforskas på 1700-talet var entomologi, och insekternas metamorfos var därför ganska okänd. I undersökningen studeras beteckningar för begreppet ’metamorfos’ jämsides med beteckningar för ägg-, larv-, pupp- och insektsstadiet. Resultaten visar att redan konventionaliserade beteckningar (*ägg*, *insekt*) ändras mycket lite i valet mellan synonyma beteckningar alternativt inte varierar alls, som ordet *ägg*. Däremot visar sig den vetenskapliga utvecklingen påverka ordförrådet så att vissa, från början starkt konventionaliserade ord såsom *mask*, successivt ersätts med beteckningen *larv* under den studerade perioden (1739–1830) (Rogström 2019, 2023).

Materialet uppvisar också lexikaliska luckor som så sakteliga fylls av nya beteckningar. Det polysema *puppa* (larven som spunnit in sig respek-

tive höljet som utgör puppans skal) utvecklas under den studerade epoken till att konventionaliseras som *puppa* respektive *kokong* – en utveckling som dock inte överensstämmer med uppgifterna i SAOB. SAOB anger ett förstabelägg från 1745 på *kokong*, men i KVAH används *kokong* inte alls förrän mot mitten av 1800-talet, och då främst i ett enda rön, Wahlberg (1848). En detaljerad studie av de polysema betydelserelationerna hos *puppa* visar att entomologerna ganska tidigt börjar skilja innehållet i puppan från dess hölje, och varierar *puppa* med en stor uppsättning synonyma uttryck, t.ex, *hylsa, hölje, skal* (Rogström 2023). Den lexikaliseringsprocess som ger sig till känna i de entomologiska rönen i KVAH hade kunnat studeras mer i detalj av SAOB:s ordboksredaktion om materialet funnits tillgängligt och uppmärkt med vetenskapsområde, varvid uppgiften om förstabelägget kunnat nyanseras något.

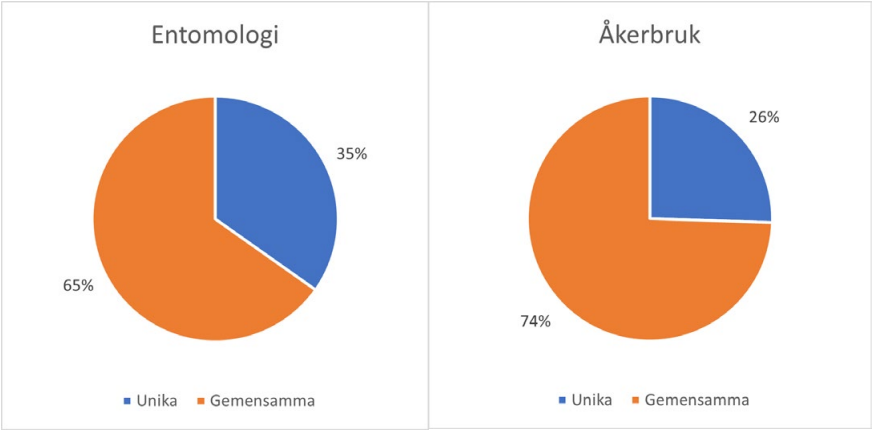
De konventionaliseringsprocesser som kan studeras för beteckningar av metamorfosen stämmer väl med de studier som genomförts av Temmerman (2000) i kartläggningen av termer inom den moderna vetenskapen *life sciences*. Detta tyder på att undersökningar av ett historiskt material kan hjälpa oss att förstå utvecklingen av ett vetenskapligt ordförråd i modernt språkbruk, något som är av värde i både modern och historisk lexikalisk forskning.

4.2. Lexikal jämförelse av ämnesområden i KVAH

Genom att märka upp materialet med uppgifter om vetenskapsområde underlättas jämförelser av lexikaliska profiler mellan texter i olika ämnen. Eftersom de enda ämnesområden som hittills är markerade hör till de manuellt inskrivna texterna om entomologi och åkerbruk genomfördes en undersökning av dessa områden.

Resultaten visar att entomologitexterna innehåller 35 % unika ord medan åkerbrukstexterna endast innehåller 26 % unika ord. Övriga ord var gemensamma inom de båda ämnena med 65 % respektive 74 % (se figur 2). Skillnaden i texternas ordförekomster kan tolkas så att de entomologiska texterna innehåller fler termer än åkerbrukstexterna, vilket kan förklaras dels med de två områdenas olika ålder, dels deras olika karaktär där entomologi har en starkare vetenskaplig koppling. Följande exempel anger unika entomologiska ord: *fjärill, koraller, kreatur.* Följande unika ord är hämtade från åkerbruk: *fruchtbar, oljach-*

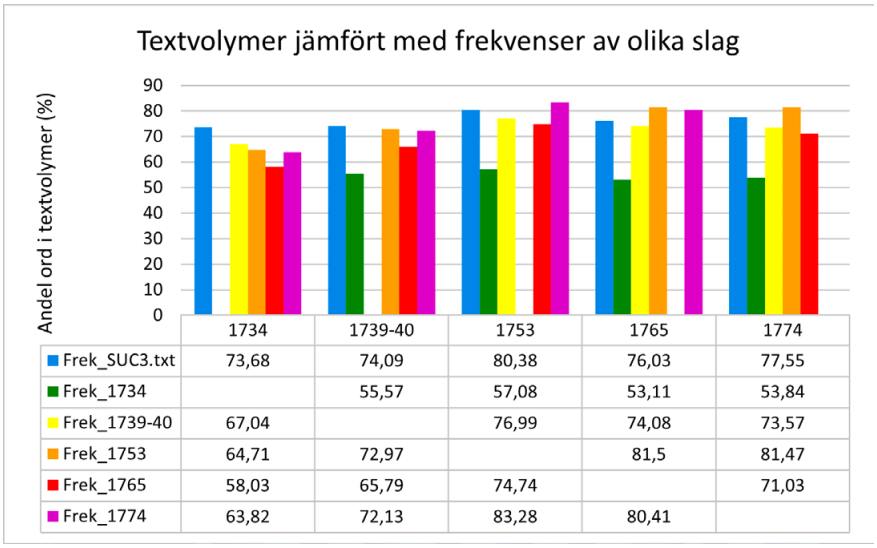
tigt, *widden*. Gemensamma ord är t.ex. följande: *hushållare*, *skapnad*, *synbar*.



FIGUR 2. Illustration av procentuell skillnad i unika och gemensamma ord för entomologi resp. åkerbruk.

4.3. Jämförelse mellan KVAH och modernt allmänspråkligt och allmänakademiskt språk

KVAH-materialet erbjuder också jämförelser med modernt ordförråd, vilket är värdefullt för studier av t.ex. betydelseutveckling. Pilotundersökning tre utgår från analyser gjorda med hjälp av Språkbankens analysverktyg Sparv i syfte att kunna upprätta lexikaliska profiler över såväl specifika ämnesområden som enskilda författare. I detta arbete är ett av våra mål att kunna visa hur delar av den ortografiska variationen kan kopplas till lexikon i Karp (Borin et al. 2012a) och därigenom förbättra analysmöjligheterna också för texter från 1700-talet. Ett annat mål är att med utgångspunkt ifrån analysen i Sparv studera vissa textuella egenskaper som t.ex. nominalkvot och ordvariation som kan vara viktiga indikationer på skillnader i informella och formella sätt att uttrycka sig. Därför har vi identifierat vardagligt, allmänakademiskt och akademiskt språkbruk. Vi har jämfört texter från fyra olika årtionden, 1740-, 1750-, 1760- samt 1770-tal, med varandra. Dessa texter har vi också jämfört med modernt språkbruk i SUC 3.0. Dessutom har jämförelser gjorts med den akademiska ordlistan.



	1734	1739-40	1753	1765	1774
Frek_SUC3.txt	73,68	74,09	80,38	76,03	77,55
Frek_1734		55,57	57,08	53,11	53,84
Frek_1739-40	67,04		76,99	74,08	73,57
Frek_1753	64,71	72,97		81,5	81,47
Frek_1765	58,03	65,79	74,74		71,03
Frek_1774	63,82	72,13	83,28	80,41	

FIGUR 3. Jämförelser av ord mellan texter från olika årtionden samt mellan modernt språkbruk och akademiska ord.

Av figur 3 framgår att överensstämmelsen mellan de äldre texterna och de moderna ordförråden i SUC 3.0 respektive den akademiska ordlistan ökar över årtiondena. Figuren visar också att det gemensamma ordförråd som används i respektive volym av KVAH blir större över tid. Andelen ord som användes 1774 har exempelvis ökat från 72 % i jämförelsen med volymerna 1739–40 till att vara 80 % i jämförelsen med 1765 års volym. Detta kan tyda på att framför allt de akademiska orden (som ju är ämnesneutrala) blir vanligare i takt med att genren utvecklas.

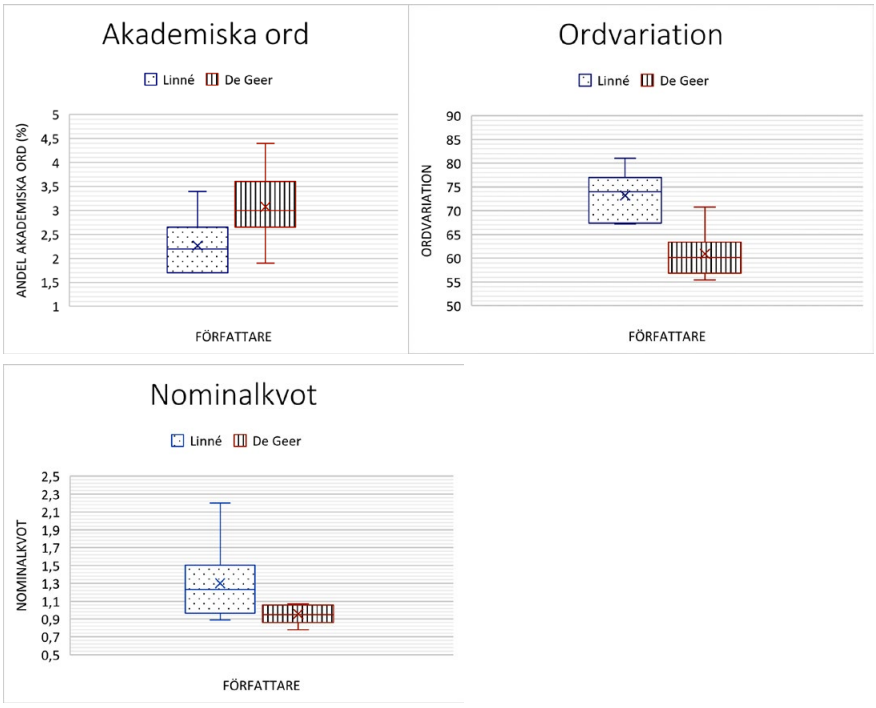
4.4. Jämförelse mellan enskilda författares lexikala profiler

Slutligen ges också ett exempel på författarspecifika studier som är möjliga att genomföra när materialet är uppmärkt med författarnamn. Här kan olika karakteristiska drag hos författarna läggas till grund för de lexikaliska profilerna, såsom formellt och informellt ordval, nominal och verbal stil, val av akademiska ord samt mer allmänspråklig variation i språket.

I undersökningen jämförs entomologiska texter av Carl von Linné och Carl de Geer. Som material används både deras presidietal och rön. Texterna är inte normaliserade vad avser ortografi. Den lexikaliska jämförelsen är baserad på morfosyntaktisk analys i form av ordklasstaggning

vilken utfördes med hjälp av Sparv. Den lexikaliska jämförelsen genomfördes i AntWordProfiler. Analysen av andel akademiska ord gjordes med hjälp av AntWordProfiler samt innehållet i den akademiska ordlistan. I den lexikaliska jämförelsen fokuseras typiska akademiska markörer, dvs. andel akademiska ord, hög nominalkvot samt ordvariation.

Resultatet visar att Linnés texter består av en mindre andel av de akademiska ord som används idag jämfört med de Geers texter. Ordvariationen hos de båda författarna visar att Linné i betydligt högre grad varierar sitt språkbruk än vad de Geer gör. Slutligen visar analysen av nominalkvot att Linné både har en högre nominalkvot samt en större variation vad gäller nominalkvot. Det innebär att hans texter är skrivna med olika nominal grad vilket i sin tur kan tolkas som att han varierar sitt språkbruk i hur formellt och informellt han skriver. De Geers texter uppvisar nominal stil som både har lägre värden än Linnés texter och dessutom är mindre varierad (se figur 4). Fries (1996:98) beskriver Linnés sätt att skriva på svenska som ”omvittnat originellt”, något som denna sista pilotstudie verkar kunna bekräfta.



FIGUR 4. Jämförelser av lexikala profiler mellan Carl von Linné och Carl de Geer.

Jämförelser av olika författares lexikala användning är intressant i sig, men kan också kasta ljus över framväxten av ett visst språkbruk relaterat till en viss författares produktionsvolym, vetenskapliga status eller annat inflytande. Listorna kan utnyttjas för excerpering av ordboksmaterial med direkt koppling till författare samt läggas till grund för studier av fackspråksetablering, vetenskapligt språkbruk och lexikalisk utveckling generellt sett.

5. Summering och framåtblick

Ur ett lexikaliskt perspektiv är 1700-talet ett dynamiskt århundrade. Svenskan erövrade allt fler domäner från latinet, nya genrer uppstod och behovet av språknormering identifierades. I ett första skede axlades detta ansvar av Kungliga Vetenskapsakademien genom att akademien använde svenska som publiceringsspråk samt understödde flera grammatik- och ordboksarbeten. En intressant fråga som lyftes redan av Gunnarsson (1987) är om författarna till rönen i KVAH eftersträvade en enhetlig vetenskaplig stil och i så fall på vilket sätt.

Ulf Teleman (2011:76) menar att terminologin för 1700-talsforskaren var ett mindre problem eftersom denne kunde förlita sig på redan existerande latinska termer – ett påstående som det finns stor anledning att granska. Redan i de fyra pilotundersökningar som redovisas i denna artikel går det att se incitament till större undersökningar som vore värdefulla för kunskapen om lexikalisk etablering under en viktig epok i vår språkliga historia. I studien av entomologiskt ordförråd (avsnitt 3.1) identifieras en lexikalisk lucka för en idag vedertaget vetenskaplig beteckning (*kokong*). Redan dessa resultat, begränsade till ett enda ämne och specifika, namngivna författare, visar att en större materialsamling skulle kunna erbjuda mer detaljerade studier över hur ordförrådet etablerades inom enskilda områden och vilka författare som förmodligen var viktigast inom vissa domäner, något som får stöd i de upprättade lexikala profilerna hos Linné och de Geer (avsnitt 4.4). Mer övergripande jämförelser av ordförråd mellan olika vetenskapliga domäner (avsnitt 4.2) samt mellan äldre och modernt akademiskt språkbruk (avsnitt 4.3) förstärker uppfattningen att denna typ av undersökningar skulle kunna bidra till lexikalisk forskning på ett flertal sätt, värdefull inte minst för historiska ordböcker. Vår kunskap om den lexikografiska utvecklingen under 1700-

talet är väl dokumenterad (se Hannesdóttir 1998) men kunskapen om det ordförråd som lexikonen skildrar kan fördjupas.

Genom större, sökbara material skulle en grundläggande kunskap om grunderna för svenskt vetenskapligt ordförråd kunna vinnas och skapa större förståelse för de processer som ligger till grund också för strukturen hos modernt fackspråkligt ordförråd. I nuläget är det främst den begränsade tillgången på material som hindrar sådana undersökningar.

Referenser

AntConc <http://www.laurenceanthony.net/software/antconc/>. Hämtad september 2022.

AntWordProfiler <https://www.laurenceanthony.net/software/antwordprofiler/>. Hämtad september 2022.

Borin, Lars, Markus Forsberg, Leif-Jöran Olsson & Jonatan Uppström 2012a. The open lexical infrastructure of Språkbanken. I: Calzolari, Nicoletta (red.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation: May 23–25, 2012*. 3598–3602.

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

Borin, Lars, Markus Forsberg, Martin Hammarstedt, Dan Rosén, Roland Schäfer & Anne Schumacher 2016. Sparv: Språkbanken's corpus annotation pipeline infrastructure. I: *The Sixth Swedish Language Technology Conference (SLTC)*, Umeå University, 17–18.

En svensk akademisk ordlista <https://spraakbanken.gu.se/ao/>. Hämtad september 2022.

Fries, Sigurd 1996. Lärdomsspråket under frihetstiden. I: Lennart Moberg & Margareta Westman (red.), *Svenskan i tusen år. Glimtar ur svenska språkets utveckling*. Stockholm: Norstedts, 88–103.

Gunnarsson, Britt-Louise 1987. Textmönster i vår äldsta vetenskapliga tidskriftsprosa. En analys av medicinska rön i Vetenskapsakademiens Handlingar 1750–1769. *Nysvenska studier* 67, 155–180.

Gunnarsson, Britt-Louise (red.) 2011a. *Languages of Science in the Eighteenth Century*. Berlin/Boston: De Gruyter Mouton.

Gunnarsson, Britt-Louise 2011b. Introduction: Languages of science in

the eighteenth century. I: Britt-Louise Gunnarsson (red.), 3–21.
Hannesdóttir, Anna Helga 1998. *Lexikografihistorisk spegel. Den enspråkiga svenska lexikografins utveckling ur den tvåspråkiga.* (Meijerbergs arkiv för svensk ordforskning 23.) Göteborg: Meijerbergs institut för svensk ordforskning.
Hellant, Anders 1753. Rön på Ebb och Flod vid Vårdhus och Nord-Caps-tracten. *Kungliga Vetenskapsakademiens Handlinga*r. Stockholm, 173–185.
Johansson, Sofie & Elisabeth Ohlsson 2019. Visualizing Vocabulary: An Investigation into Student Assignments in CLIL and Non-CLIL Contexts. I: *Investigating Content and Language Integrated Learning, Multilingual Matters*, 216–235.
Johansson, Sofie & Clas, Olander 2022. Ämneslitteracitet inom skolans naturvetenskap. I: Jakobsson, Anders, Pia Nygård Larsson & Lotta Bergman (red), *Ämneslitteracitet och inkluderande undervisning*. Lund: Studentlitteratur, 179–212.
Karp <https://spraakbanken.gu.se/verktyg/karp>. Hämtad september 2022.
Korp <https://spraakbanken.gu.se/verktyg/korp>. Hämtad september 2022.
KVAH = Kungliga Vetenskapsakademiens Handlingar <https://hosting.devo.se/kvah/search.html>. Hämtad september 2022.
Landqvist, Hans, Lena Rogström & Greta Horn (2020). ”Dock medgifver jag, at af et års Rön ej kan göras någon allmän slutsats.” Clas Bjerkander konstruerar och positionerar sig som vetenskaplig skribent. I: Daniel Sävborg, Eva-Liina Asu & Anu Laanemets (red.), *Språkmöte och språkhistoria. Studier i svensk språkhistoria 15*. (Nordistica Tartuensia 21.) Tartu: University of Tartu Press, 148–160.
Lindroth, Sten 1967. *Kungl. Svenska Vetenskapsakademiens Historia 1739–1818. II Tiden 1783–1818*. Stockholm: Kungl. Vetenskapsakademien.
Lundbladh, Carl-Erik 1992. *Handledning till Svenska Akademiens Ordbok*. Stockholm: Norstedts.
Nilsson, Pär 2019. *Bildliga betydelser i SAOB. Om beskrivningen av betydelseutvecklingsmekanismer analyserad ur ett kognitivt semantiskt perspektiv*. (Lundastudier i nordisk språkvetenskap A 79.) Lund: Språk- och litteraturcentrum, Lunds universitet.

Rogström, Lena 2019. *Fackordförråd i Kungl. Vetenskapsakademiens Handlingar. En pilotstudie med fokus på entomologi 1739–1854*. GU-ISS-2019-01. Göteborg: Forskningsrapporter från institutionen för svenska språket, Göteborgs universitet.

Rogström, Lena 2023. Ord för det okända. Om lexikalisk utveckling i 1700-talsentomologi. I: Lars-Olof Delsing & Bo-A. Wendt (red.), *Studier i svensk språkhistoria 16. Främmande inflytande på svenska språket.* (Lundastudier i nordisk språkvetenskap A 85.) Lund: Språk- och litteraturcentrum, Lunds universitet, 151–160.

SAOB= *Ordbok över svenska språket utgiven av Svenska Akademien*. 1898–. Lund: Gleerups.

SAOB <https://svenska.se/>. Hämtad september 2022.

Santesson, Lillemor 1986. *Tryckt hos Salvius. En undersökning om språkvården på ett 1700-talstryckeri med särskild hänsyn till ortografi och morfologi.* (Lundastudier i nordisk språkvetenskap A 37.) Lund: Lund University Press.

Sköldberg, Emma & Sofie Johansson Kokkinakis 2012. A och O om akademiska ord. Om framtagning av en svensk akademisk ordlista. I: Birgit Eaker, Lennart Larsson & Anki Mattisson (red.), *Nordiska studier i lexikografi* 11. Lund, 575–585.

Sparv <https://spraakbanken.gu.se/verktyg/sparv>. Hämtad september 2022.

SUC 3.0 = Stockholm-Umeå-Corpus 3.0 <https://spraakbanken.gu.se/resurser/suc3>. Hämtad september 2022.

Teleman, Ulf 2011. The Swedish Academy of Science: language policy and language practice. I: Britt-Louise Gunnarsson (red.), 63–87.

Temmerman, Rita 2000. *Towards new ways of terminology description: the socio-cognitive approach*. Amsterdam, Philadelphia PA: John Benjamins.

Transkribus <https://readcoop.eu>. Hämtad september 2022.

Wahlberg, Peter Fredrik 1848. Ytterligare bidrag till kännedom om Svamp-myggan Ceroplatus sesioides. *Kungliga Vetenskapsakademiens Handlingar*. Stockholm, 317–327.

Nyord i to norske ordbøker

Dagfinn Rødningen & Knut E. Karlsen

This paper examines a selection of neologisms from the two general dictionaries *Bokmålsordboka* and *Nynorskordboka*. The words have been quantified and analysed based on morphological, semantic and orthographical factors. The results show that neologisms of Norwegian origin constitute 69% of the material, while neologisms of English origin constitute 22% of the material. The vast majority of neologisms of Norwegian origin are compounds, while simple neologisms constitute a larger part of the words of foreign origin. Semantically, 39% of the words of foreign origin fit into a category of words dealing with internationalization, whereas only a few of the neologisms of Norwegian origin fit into this category. An analysis shows that 49% of all neologisms of English origin in the material have been given a spelling identical to the English original, while 43% have a spelling more or less in accordance with principles of Norwegian orthography and pronunciation. Only a small group of words have been given both English and Norwegian spelling.

NØKKELORD: nyord, norsk opphav, importord, semantiske kategorier, skrivemåte

1. Innledning

Når språk utvikler og fornyer seg, er dannelse av nyord en viktig del av prosessen. Kartlegging og analyse av nyord er dermed vesentlig i observasjon av språkutvikling. Vi som har skrevet denne artikkelen, jobber i Språkrådet i Norge. En sentral oppgave for Språkrådet er å normere de norske skriftspråka, bokmål og nynorsk, og som et ledd i dette arbeidet skal vi observere språkutviklingen, deriblant innslag av nye ord i norsk. For å få et inntrykk av opptaket av nyord de siste åra har vi tatt for oss to norske allmennordbøker, Bokmålsordboka og Nynorskordboka, med sikte på å trekke ut nyord derfra som grunnlag for en analyse av nyordsdannelsen i norsk.

Fra omkring midten av forrige århundre samarbeidet de nordiske språknemndene om innsamling og kartlegging av nyord. Målet var blant annet å få fram et felles grunnlag for nyordsbøker på de skandinaviske språka (Norsk språkråd 1982:9–10).

For norsk resulterte samarbeidet i den første av to nyordsbøker (Norsk språkråd 1982). Utgiver var Språkrådets forgjenger, Norsk språkråd.

Boka inneholder mellom 7000 og 8000 nyord i norsk fra perioden 1945–1975. Orda er stort sett hentet fra seddelsamlinger basert på skriftlige kilder som aviser og tidsskrifter.

I 2012 kom en ny ordbok med nyord i norsk (Guttu & Wangensteen 2012), denne gang med nyord fra perioden 1975 til 2005. Arbeidet ble utført av to leksikografer ved Universitetet i Oslo (UiO) etter initiativ fra Språkrådet. De fleste av de ca. 10 000 oppslagsorda er hentet fra nyordsmaterialet ved UiO (1968–1998) og fra Språkrådets nyordssamling (1975–2001). Dette er tradisjonelle seddelsamlinger. Den nyeste delen av materialet er hentet fra Norsk aviskorpus ved Universitetet i Bergen (UiB), og det metodiske grunnlaget for deler av boka er dermed noe annerledes og mer moderne enn for den første nyordsboka. I tillegg til oppslagsord med definisjoner og eksempler inneholder begge ordbøkene beskrivelser og analyser av nyordsmaterialet.

Det er ikke gjennomført noen helhetlig kartlegging av nyord i norsk etter 2005. Språkrådet registrerer nyord som del av vårt generelle arbeid med språkobservasjon. Vi gjennomfører dessuten en årlig kåring av årets ord (se Andersen & Våge 2014), men i motsetning til for eksempel Dansk Sprognævn har ikke Språkrådet i sitt mandat å drive systematisk innsamling av nyord eller gi ut ordbøker med nyord i norsk.

2. Mål, metode og materiale

2.1. Kategorisering

Målet med undersøkelsen har vært å kvantifisere et utvalg nyord i norsk i nyere tid og gjøre greie for visse morfologiske, semantiske og ortografiske kjennetegn ved nyorda. Vi har ønsket å tallfeste fordelingen blant nyorda mellom ord med engelsk opphav, ord med opphav fra andre språk og ord med norsk opphav, blant annet for å teste en hypotese om at sammensatte nyord i hovedsak baserer seg på hjemlig norsk språkmateriale, mens simpleks blant nyorda, altså ord som hverken er sammensatt eller avledet, for en stor del utgjøres av importord. Videre presenterer vi noen semantiske kategorier som er identifisert i arbeidet med nyorda, og ser nærmere på hvilke kategorier som preges av importord, og hvilke som preges av sammensetninger med norsk opphav. Når det gjelder rettskrivingspraksis for importord, har vi sett på graden av tilpasning til norske rettskrivingsregler.

2.1.1. Morfologi

Vi opererer med et grunnleggende skille mellom nyord av norsk opphav og nyord som er importord. Importord er et samlebegrep for fremmedord og lånord og utgjør dermed den delen av ordforrådet som ikke er arveord. Forskjellen er at fremmedord ikke er tilpasset norsk språkstruktur, mens lånorda er det (Sandøy 2000:19). Siden graden av tilpasning til norske rettskrivingsregler er en del av denne undersøkelsen, er det nettopp et poeng å bruke overbegrepet importord. Vi skiller også mellom sammensatte og avledede ord på den ene siden og simpleks, kortformer og forkortelser på den andre. Vi har forsøkt å holde blandingsformer utenfor, for eksempel sammensatte ord hvor vi har ledd av både norsk og fremmed opphav. Blant de sammensatte orda kan det forekomme tilfeller der ett eller flere ledd er ord med opprinnelig fremmed opphav, men hvor ordet er blitt så innarbeidet i norsk at vi i denne sammenhengen ikke lenger ser på det som et importord. Vi har regnet oversettelseslån som egen kategori. Oversettelseslån er ord som mer eller mindre er direkte norske oversettelser av importord, og de blir i denne sammenhengen regnet som norske nyord.

Blant importorda skiller vi mellom importord fra engelsk og importord fra andre språk. I kategorien engelske importord finner vi ord som er lånt fra både britisk og amerikansk engelsk, i tillegg til ord som er lånt inn fra andre språk via enten britisk eller amerikansk engelsk.

2.1.2. Semantikk

Vi har definert tre semantiske hovedkategorier som grovt sett dekker store deler av materialet. Disse kategoriene er

- sosial, økonomisk og vitenskapelig utvikling
- internasjonalisering
- teknologisk utvikling og nyvinning

Det kan være glidende overganger mellom kategoriene, og flere av orda i materialet kan være vanskelige å plassere. Her har vi måttet basere oss på skjønn. Videre vil særlig den første kategorien inneholde semantisk svært ulike ord, som godt kunne dannet grunnlag for ytterligere kategorisering

i undergrupper. Vi finner ord knyttet til så ulike semantiske områder som for eksempel medisin og sykdom, endrede familiestrukturer, geografiske og administrative endringer og kjønnsnøytrale termer. For vårt hovedformål har vi likevel ikke sett det som hensiktsmessig å gjøre en mer finmasket inndeling enn de tre hovedkategoriene.

2.1.3. Ortografi

I analysen av rettskrivingspraksis for importorda har vi avgrenset materialet til å gjelde engelske importord. Vi har undersøkt i hvilken grad importorda får fornorsket skrivemåte, ved å skille mellom ord som beholder original engelsk skrivemåte som eneform, og ord som i større eller mindre grad er tilpasset norsk rettskriving eller uttale. I tillegg kommer en gruppe ord der det er valgfrihet mellom opprinnelig engelsk skrivemåte og tilpasset norsk skrivemåte.

2.2. Kilder og grunnlagsmateriale

2.2.1. *Bokmålsordboka* og *Nynorskordboka*

Vi har valgt å hente materialet for undersøkelsen vår fra *Bokmålsordboka* (BOB) og *Nynorskordboka* (NOB), rettskrivings- og definisjonsordbøker som er eid av Språkrådet og UiB i fellesskap. De inneholder ord fra det allmenne ordforrådet og skal vise hva som til enhver tid er korrekt staving og bøying av oppslagsorda som er tatt med. Ordbøkene kom ut i 1986, med ca. 90 000 ord i NOB og 60 000 ord i BOB. En nettversjon ble publisert første gang i 1994. Siste papirversjoner ble utgitt i 2005 (BOB) og 2012 (NOB). Senere endringer og utvidelser har kun blitt publisert i nettversjonen av ordbøkene.

Ved å bruke BOB og NOB som kilder for undersøkelsen kan vi være sikre på at den registrerte skrivemåten av orda er i tråd med den offisielle rettskrivingen. Det gir større troverdighet til resultatene av rettskrivingsdelen av undersøkelsen.

2.2.2. Ordmaterialet

Vi har tatt utgangspunkt i nyopprettede artikler i nettutgaven av ordbøkene. Av tekniske årsaker har det ikke vært mulig å hente artikler fra før

2008. Bestillingen vår til teknikerne ved UiB var dermed å gjøre et uttrekk av alle nye artikler opprettet etter 2008. Materialet går fram til 2019, da pandemistengingen gjorde at vi måtte sette stopp for innsamlingen. Uttrekket gav oss to rålister, én for NOB og én for BOB, med til sammen 10 170 ord.

Listene inneholder mye mer enn nyord. Majoriteten av de nyopprettede artiklene er tilfeldige lakuner, sammensetninger som fantes før, men ikke har kommet med tidligere av plasshensyn osv. Vi har derfor måttet gjøre en omfattende vaskejobb manuelt, men vi har hatt definerte kriterier å gå etter. Orda i grunnlagsmaterialet vårt skal for eksempel ikke ha vært nevneverdig brukt før ordbøkene ble utgitt første gang. I praksis har vi operert med en tidsgrense rundt 1980. Metodisk er det viktig å understreke dette, og vi kan ikke underslå at en viktig faktor har vært vår kjennskap til språket og samfunnet.

Om orda i grunnlagsmaterialet er dekt i de to tidligere nyordsbøkene, har ikke vært et kriterium for utvelgelse. Innholdet i Guttu & Wangensteen (2012), nyord fram til 2005, vil dels ha blitt tatt opp i allmennordbøkene senere, dels ha falt ut av allmenn språkbruk og dermed ikke oppfylle kriteriene for opptak i allmennordbøkene. Vår primære interesse er nyord som har blitt en del av allmennspråket, og det kan ha skjedd etter 2005 for mange av orda vi finner i Guttu & Wangensteen (2012).

Vi skal gjøre oppmerksom på at en stor del av nyorda i BOB og NOB har fått opprettet artikkel i siste del av undersøkelsesperioden, særlig fra 2017 og utover. Det har sammenheng med at forberedelsene til det pågående revisjonsprosjektet for NOB og BOB startet da. De nye artiklene fra 2017 og utover ligger i første delen av alfabetet. I revisjonen oppdaterer redaktørene systematisk lemmautvalget, og når revisjonen er ferdig, vil antallet nyord være betraktelig høyere enn det vi presenterer her. Derfor legger vi ikke særlig vekt på kvantitative forhold i undersøkelsen. Vi tror likevel at tendensen vi ser i denne delen av materialet, også vil gjelde for ordbøkene i sin helhet.

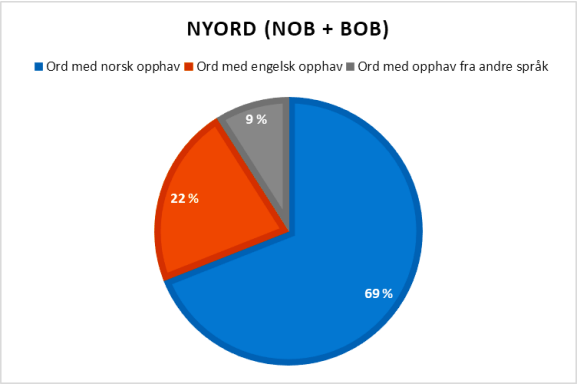
Når listene var rensket og flettet og vi hadde fjernet dubletter, stod vi igjen med til sammen 335 unike nyord. Det er disse nyordene som utgjør empirien – datagrunnlaget – for denne undersøkelsen.

3. Funn

3.1. Morfologi

3.1.1. Nyord av norsk opphav vs. importord

Som det går fram av figur 1, er 69 prosent (230 ord) av nyorda i materialet vårt av norsk opphav, mens 31 prosent (105 ord) er importord. Engelske importord utgjør den største delen av importorda og 22 prosent (74 ord) av hele materialet. 9 prosent (31 ord) av nyorda i materialet er lån fra andre språk enn engelsk. De største lånspråka ved siden av engelsk er arabisk, spansk, fransk, japansk, italiensk, tyrkisk og latin (rangert i fallende rekkefølge).



FIGUR 1. Nyord fordelt etter etymologisk opphav.

Blant nyorda av norsk opphav finner vi en gruppe nyord (16 ord) som bare gjelder nynorsk, nemlig lån fra bokmål. I Norge er det staten som normerer skriftspråket, men det gjelder bare bøying og skrivemåte. Ordtilfanget er ikke normert, men opptak av danske og tyske ord i nynorsk (gjennom bokmål) er et viktig unntak. I dag er frekvens en viktig faktor i normering av skriftspråka, og dette hensynet kommer ofte i konflikt med ønsket om å holde ord av dansk og tysk opphav ute fra nynorsk. Det er derfor en økende tendens til at ord som disse finner veien inn i NOB: *antal* (tradisjonelt *talet på*), *menigheit* (tradisjonelt *kyrkjelyd*), *beskyttelse* (tradisjonelt *vern, hjelp, verje*), *bekymra* (tradisjonelt *uroa*), *midlertidig* (tradisjonelt *førebels*), *bevisstlaus* (tradisjonelt *medvitslaus*). Denne kategorien er en viktig forklaring på at tallet på nyord er noe høyere i NOB enn i BOB.

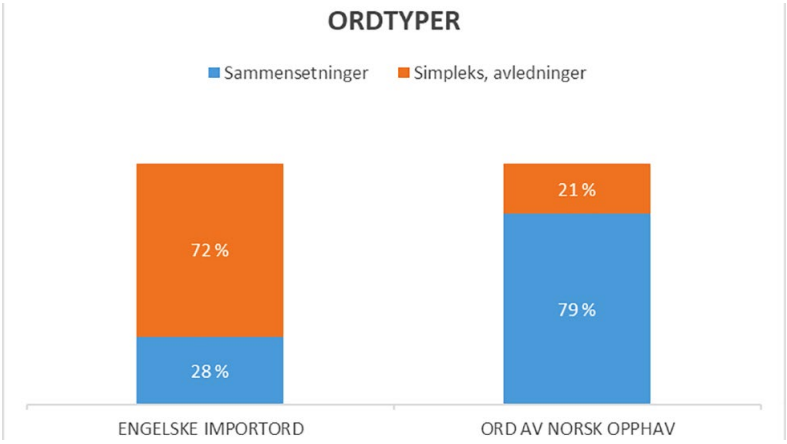
Tabell 1 viser eksempler på nyord av ulike opphav. Orda i tabellen er tatt opp både i BOB og NOB.

TABELL 1. Eksempler på nyord av norsk opphav og nyord fra andre språk

Nyord av norsk opphav	**Nyord av engelsk opphav**	**Nyord med opphav fra andre språk**
allværsjakke	*caste*	*djembe* (mande)
berøringsskjerm	*beatboxing*	*aioli* (fransk)
KMI	*bugg/bøgg*	*bulgur* (tyrkisk)
antiskrens	*blender*	*børek* (tyrkisk)
asylbarn	*chatte*	*karaoke* (japansk)
framsnakke	*cover*	*spa* (etter belgisk kursted)
helsesjukepleiar	*cupcake*	*chorizo* (spansk)
kroppskamera	*google*	*focaccia* (italiensk)
ballbinge	*pride*	*abaya* (arabisk)
lenkeråte	*skate*	*nikab* (arabisk)

3.1.2. Ordtyper

Nyorda av norsk opphav skiller seg tydelig fra importorda når vi deler dem inn i ordtyper. I figur 2 ser vi at sammensetninger utgjør hele 79 prosent (181 ord) av nyorda av norsk opphav. Til sammenligning er andelen sammensetninger blant importord av engelsk opphav bare 28 prosent (21 av 74 ord).



FIGUR 2. Fordeling mellom sammensetninger og usammensatte ord blant nyord av norsk opphav og engelske importord.

De fleste sammensatte orda av norsk opphav er av typen substantiv + substantiv, eksempelvis *asylbarn*, *ballbinge* og *lenkeråte*. Oversettelseslån som *berøringsskjerm*, *allværsjakke* og *kroppskamera* er gjerne også av denne typen. I den andelen som ikke er sammensetninger, finner vi forkortelser som *KMI* (kroppsmasseindeks, som også er et oversettelseslån), *MR* (magnetisk resonans), *AFP* (avtalefestet pensjon) og avledninger som *aktivitør*, *antiskrens* og *biodiesel*.

I den store andelen engelske importord som ikke er sammensatte, finner vi i tillegg til kortord, forkortelser og avledninger også mange simpleks. Vi har ikke funnet simpleks blant orda av norsk opphav. Simpleks ser dermed ut til å være vanligere blant nyord av engelsk opphav enn blant nyord av norsk opphav. Dette bekrefter vår hypotese om at sammensatte nyord i hovedsak er av norsk opphav, mens simpleks blant nyorda for en stor del utgjøres av importord. Det er også rimelig i lys av at sammensatte ord er mindre vanlig i engelsk enn i norsk.

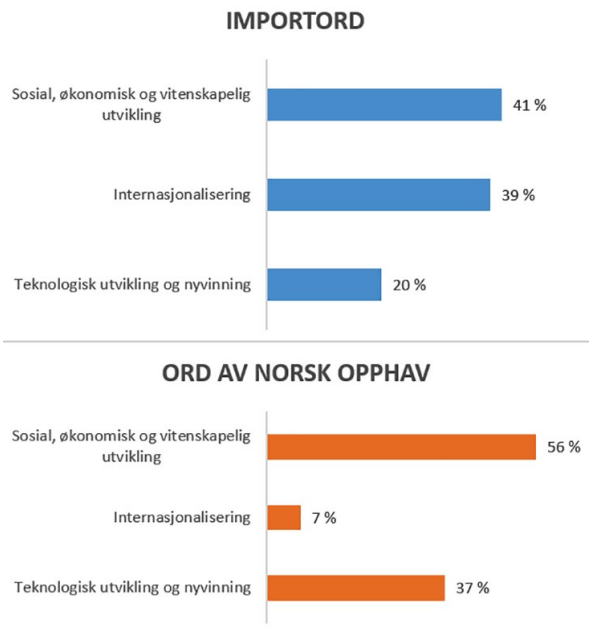
3.2. Semantikk

Fordelingen mellom importord og nyord av norsk opphav innenfor de tre semantiske hovedkategoriene[1] er som vist i figur 3 nedenfor.

Kategorien sosial, økonomisk og vitenskapelig utvikling domineres av ord av norsk opphav, som *eldrebølge*, *bilkollektiv*, *bonusbarn*, *helsesykepleier*, *musesykje* og *kortreist*. Av importord i denne kategorien finner vi for eksempel *disse* ('snakke stygt om'), *emoji* og *babyshower*.

I internasjonaliseringskategorien er det importorda som dominerer. Vi finner ord av engelsk opphav, som *cheeseburger*, *halloween* og *backpacker*, men her er også andre långiverspråk godt representert: i ord som har med mat og drikke å gjøre (*cava*, *cru*, *falafel*, *chorizo*), ord med tilknytning til religion (*nikab*) og betegnelser for folk og kulturer (*saharawi*). Nyord av norsk opphav er det ikke mange av i denne kategorien, men vi finner et eksempel som *europolitiker* og ord som oppstår når nye stater dannes, som *sørsudaner*.

1 I inndelingen i semantiske kategorier er gruppa «bokmålslån i nynorsknormen» holdt utenfor datagrunnlaget.



FIGUR 3. Fordeling mellom importord og ord av norsk opphav etter semantisk kategori.

At 39 prosent (41 ord) av importorda havner i kategorien internasjonalisering, er ikke overraskende. Det er kanskje også rimelig at så mange som 56 prosent (118 ord) av nyorda av norsk opphav befinner seg innenfor en såpass vid semantisk kategori som sosial, økonomisk og vitenskapelig utvikling. Når vi ser på kategorien teknologisk utvikling og nyvinning, er imidlertid fordelingen mellom importord og nyord av norsk opphav jevnere, selv om orda av norsk opphav dominerer også der. Tabell 2 viser noen eksempler på nyord fordelt på semantiske kategorier. Eksempler på ord av norsk opphav i kategorien «teknologisk utvikling og nyvinning» er *nettbrett*, *betalingsmur*, *fellski* og *smarttelefon*. Det er verdt å merke seg at alle importord i denne kategorien kommer fra engelsk, og at ord tilknyttet IT-utvikling dominerer. Typiske ord her er *app*, *google*, *LED/led* og *ruter*.

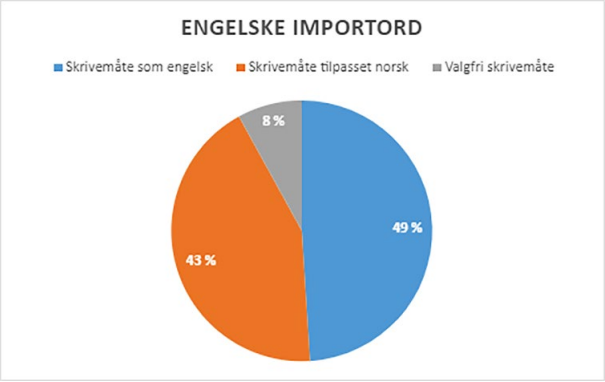
TABELL 2. Eksempler på nyord (importord og ord av norsk opphav) fordelt på semantiske kategorier

	Sosial, økonomisk og vitenskapelig utvikling	**Internasjonalisering**	**Teknologisk utvikling og nyvinning**
Importord	*blogger* *cheerleader* *cisperson* *emotikon* *emoji* *fantasy* *hivsmitte* *pride* *skateboard*	*abaya* *backpacker* *balsamico* *cava* *cru* *falafel* *halloween* *nikab* *saharawi*	*app* *backup* *bugg/bøgg* *batch* *DAB* *DVD* *google* *LED/led* *ruter*
Ord av norsk opphav	*bemanningsbyrå* *bilkollektiv* *bonusbarn* *fastlege* *medmor* *klimautslipp* *nettauksjon* *språkbad*	*abkhasisk* *abkhasar* *eswatiniar* *eswatinisk* *europolitiker* *nordmakedonsk* *sørsudaner* *sørsudansk*	*alkolås* *betalingsmur* *elsykkel* *fellski* *nettbrett* *skjermdump* *smarttelefon* *trykkskjerm*

3.3. Ortografiske tilpasninger

Vi har ønsket å se på hvordan importorda blir normert i norsk. Dette har vært mulig fordi vi har brukt normative ordbøker som kilder for undersøkelsen. Av totalt 105 importord i materialet er 70 prosent (74 ord) av engelsk opphav. Det er disse 74 engelske importorda som er grunnlaget for denne delen av undersøkelsen, og de fordeler seg som vist i figur 4:

FIGUR 4. Skrivemåten av engelske importord.



3.3.1. Original engelsk skrivemåte som eneform

49 prosent (36 ord) er normert med den engelske opphavsformen som eneform. Eksempler på slike ord er a*udition*, *beatboxing*, *coach* og *cupcake*. I normering av norsk har det vært tradisjon for å forsøke å fornorske stavemåten av innlånte ord dersom man ikke har funnet et norsk alternativt ord eller et godt oversettelseslån (jf. Sandøy 2000:211–212 og Språkrådet 2021, pkt. 10.1). Erfaringen viser likevel at det kan være vanskelig å få gjennomslag for fornorskede skrivemåter dersom den fremmede skrivemåten har fått etablere seg i bruk innen ordet blir tatt opp i ordbøkene og dermed får en normert form. De nyere prinsippene for normering av norsk legger mer vekt på faktisk språkbruk (usus) enn tidligere. Det kan være en forklaring på at nesten halvparten av importorda ikke har fått fornorsket form.

3.3.2. Større eller mindre tilpasning til norsk rettskriving og uttale

43 prosent (32 ord) av de engelske importorda har fått skrivemåte som er mer eller mindre tilpasset norske rettskrivingsregler eller norsk uttale. Når ord normeres med fornorsket skrivemåte, har det ofte allerede oppstått en slik skrivemåte blant brukerne, enten for det spesifikke ordet eller for ord med lignende kjennetegn. Et gjennomgående trekk er at verb får lagt til norsk infinitivsending, som i verbet *chatte*. I noen tilfeller får uttalen forrang i rettskrivingen når den skiller seg tydelig fra den fremmede skrivemå-

ten, som i adjektivet *døll* for engelsk *dull*. Visse bokstavsamband kan gjengis på en standardisert måte. Slik kan *-ck-* få norsk skrivemåte *-kk-*, som i *dokkingstasjon*. Vanlige ordelementer som allerede finnes i norske importord, kan overføres til nye importord. Da kan vi få skrivemåter som *casestudie*.

3.3.3. Valgfrihet mellom opprinnelig skrivemåte og tilpasset norsk skrivemåte

I noen tilfeller velger man å la en fornorsket skrivemåte stå som alternativ form til den opprinnelige skrivemåten av importord. Slik får språkbrukerne et valg, og man kan etter noen år eventuelt endre normeringen dersom man ser at det ene eller andre alternativet blir dominerende. Bare et fåtall (8 prosent (6 ord)) av de engelske importorda i materialet vårt er normert på denne måten. Eksempler på slike ord er *quilt* eller *kvilt* (opprinnelig *quilt*), *RIB* eller *ribb* (opprinnelig *RIB*), *server* eller *sørver* (opprinnelig *server*) og *LED* eller *led* (opprinnelig *LED*).

4. Oppsummering og konklusjon

Analysen av de 335 nyorda som utgjør empirien i denne undersøkelsen, viser at flertallet av nyorda er av norsk opphav (69 prosent). Av nyorda av norsk opphav er 79 prosent sammensetninger. Blant importorda dominerer ord av engelsk opphav. 70 prosent av orda i denne kategorien er ord fra engelsk. Eksempler på andre långiverspråk er arabisk, spansk, fransk, japansk, italiensk, tyrkisk og latin.

Nyord av norsk opphav er vanligst i den semantiske kategorien vi har kalt *sosial, økonomisk og vitenskapelig utvikling*, mens importorda dominerer klart i den semantiske kategorien *internasjonalisering*. Her står de engelske importorda sterkt, men også andre långiverspråk er representert, særlig når det gjelder ord som har med mat og drikke å gjøre.

49 prosent av engelske importord har uendret skrivemåte, mens 43 prosent av de engelske importorda har skrivemåte tilpasset norsk. En liten gruppe ord (8 prosent) er normert med både engelsk og norsk skrivemåte. At nesten halvparten av de engelske importorda har blitt normert med original engelsk skrivemåte, kan tyde på at man i begrenset grad har innfridd ønsket om tilpasset norsk skrivemåte av importord. De gjeldende retningslinjene for normering av importord (Språkrådet 2021, pkt. 10)

viser at ambisjonen om tilpasning fortsatt står ved lag. Det gjenstår å se om resultatet fra undersøkelsen vår på dette punktet kan få betydning for hvordan importord i norsk blir normert i framtida. Resultatene kan også være av interesse som sammenligningsgrunnlag for hvordan skrivemåten av importord blir tilpasset i de øvrige språka i Norden.

Empirien i denne undersøkelsen omfatter nyord i NOB og BOB fram til 2019. Gjennom det pågående revisjonsprosjektet ved Universitetet i Bergen blir også lemmautvalget vurdert. Det forventes at en god del nyord vil komme til ettersom nye alfabetbolker blir revidert. Selv om vi mener å ha fanget opp viktige tendenser ved nyordsutviklinga i norsk i denne undersøkelsen, bør den følges opp av en ny etter revisjonsslutt.

Litteratur

Andersen, Gisle & Ole Våge 2014. Nyord, kriterier og språknormering i en årlig kåring. I: Andersen, Margrethe Heidemann, Pia Jarvad & Jørgen Nørby Jensen (red.), *Neologismer. Dansk Sprognævns 2. seminar om nye ord*. København: Dansk Sprognævn, 23–42.

BOB = *Bokmålsordboka*. Språkrådet og Universitetet i Bergen. <http://ordbøkene.no>. Hentet mars 2022.

Guttu, Tor & Boye Wangensteen 2012. *Nyord i norsk*. Oslo: Kunnskapsforlaget.

NOB = *Nynorskordboka*. Språkrådet og Universitetet i Bergen. <http://ordbøkene.no>. Hentet mars 2022.

Norsk språkråd 1982. *Nyord i norsk. 1945–1975*. Bergen: Universitetsforlaget.

Sandøy, Helge 2000. *Lånte fjører eller bunad? Om importord i norsk*. Oslo: Landslaget for norskundervisning og Cappelen akademisk forlag.

Språkrådet 2021. *Retningslinjer for normering av bokmål og nynorsk*. <https://www.sprakradet.no/globalassets/spraka-vare/norsk/retningslinjer-for-normering-2021--bokmalsversjon.pdf>. Hentet mars 2023.

Stærke participier i attributiv stilling – en leksikografisk udfordring

Jørgen Schack & Eva Skafte Jensen

Retskrivningsordbogen, the official dictionary of Danish orthography, contains large amounts of material inherited from previous editions, among other things, information on the gender inflection of the perfect participle for a large number of strong verbs. In the attributive position, such participles end in *-n* or *-t* in front of common gender nouns and in *-t* in front of neuter nouns. In this paper, we first account for the official rules and for the actual use of the *-n-* and *-t*-forms in Modern Danish. The situation in Modern Danish is quite diverse, and many language users are unsure of what to do with the gender inflection of the participles. Next, we present a number of lexicographic challenges resulting from this situation and provide some tentative suggestions on how *Retskrivningsordbogen* can become a bit clearer on this point.

NØGLEORD: perfektum participium, stærke verber, genusbøjning, Retskrivningsordbogen

1. Indledning

Ligesom alle andre ordbøger som bygger på tidligere udgaver, indeholder Retskrivningsordbogen store mængder overleveret stof, først og fremmest opslagsord med tilhørende bøjningsoplysninger. En af de ting som Retskrivningsordbogen har arvet fra tidligere udgaver, er oplysning om genusbøjning af perfektum participium i attributiv stilling ved et stort antal stærke verber, fx

> **forvride** *vb.*, -r, forvred, forvredet *(foran fælleskønsord* forvreden *el.* forvredet*)*, forvredne *(jf. § 31-34)*; en forvreden el. forvredet ankel; et forvredet knæ; forvredne led

Den seneste udgave af Retskrivningsordbogen (4. udg., 2012) indeholder 300 verber hvis participier har mulighed for en genusbøjet form på *-n* i attributiv stilling, fx *en forvreden ankel*. Den konkurrerende form på *-t*, fx *en forvredet ankel*, kan ifølge ordbogen – og nutidige grammatik-

ker (Hansen & Heltoft 2011:671) og sprogrigtighedshåndbøger (Galberg Jacobsen & Stray Jørgensen 2013:487) – altid vikariere for *-n*-formen, og den kan derfor betragtes som en defaultform, som kan bruges uanset substantivets genus (jf. afsnit 5.1).

De participiale former på *-n* kan ikke formelt holdes ude fra utrumsformerne af et stort antal adjektiver som har et modsvarende verbum i Retskrivningsordbogen, fx *overdreven* (adj.) : *overdrive*/*overdreven* (vb.), *stjålen* (adj.) : *stjæle*/*stjålen* (vb.). Adjektiver af denne type har i en del tilfælde neutrumsformer på *-nt*, og nogle af dem har en overført betydning, som afviger fra participiernes konkrete (verbale) betydning, fx *et stjålent blik* 'et skjult blik'. De modsvarende participiale former har derimod altid neutrumsformer på *-t*, fx *stjålet*. Der opstår derfor let usikkerhed mht. brugen af neutrumsformerne: Kan det kun hedde *et stjålet maleri*, eller kan det også hedde *et stjålent maleri*?

I artiklen gør vi først rede for reglerne for og brugen af *-n-* og *-t*-former i moderne dansk (med en meget kort historisk afstikker). Situationen i moderne dansk er ganske broget, og mange sprogbrugere er usikre på hvad de skal stille op med participiernes genusbøjning. Dernæst fremlægger vi en række leksikografiske udfordringer som er affødt af denne situation og giver nogle forsigtige bud på hvordan Retskrivningsordbogen kan blive lidt klarere på dette punkt.

2. Moderne standarddansk

Situationen i moderne standarddansk kan fremstilles på følgende måde:

	utrum	neutrum
1. *-n* og *-t* som genusbøjning	*-n*	*-t*
2. *-t* som ubøjet form/defaultform	*-t*	
3. *-n* som del af stammen (i afledte adjektiver)	*-Ø*	*-n-t*

Ad 1: Formen på *-n* er altid utrum, fx *en frosse**n** kylling*. Formen på *-t* kan være paradigmets neutrumsform, fx *et frosse**t** vandrør*.

Ad 2: Formen på *-t* kan desuden være en ubøjet form/defaultform, fx *en frosse**t** kylling* (utrum), *et frosset vandrør* (neutrum).

Ad 3: I de tilfælde hvor der ved siden af verbets participiumsform foreligger et afledt adjektiv på *-n*, må *-n* beskrives som en del af adjektivets stamme, og utrumsformen ender derfor på *-Ø*; neutrumsformen føjer *-t* til stammens *-n*, fx *en frossen-Ø kylling*, *et frossen**t** vandrør*. For overskuelighedens skyld (og af hensyn til traditionen i ordbøger og grammatikker) noterer vi det i denne artikel tilfælde på følgende måde: *en frosse**n** kylling*, *et frosse**nt** vandrør*. I et tilfælde som fx *en frosse**n** kylling* kan det altså ikke afgøres om *frosse**n*** er participiets utrumsform (som står i modsætning til *frosse**t***), eller om det er adjektivets utrumsform (som står i modsætning til *frosse**nt***). Kun former på *-nt* er entydigt adjektiviske.

3. En ultrakort historik

Dansk har oprindelig haft gennemført genusbøjning af stærke participier i attributiv stilling (ubestemt form singularis), fx *en urolig og sønderreven Sjæl* : *et sønderrevet Hjerte* (B.S. Ingemann, 1826). Afledte verbaladjektiver med *-n* som en del af stammen og heraf følgende neutrumsformer på *-nt*, dvs. former som fx *forfaldent* og *tvungent*, kendes i hvert fald fra omkring 1800 (se ODS, I. **forfalde**, *v.* og III. **tvinge**, *v.*).

Brugen af *-t*-formen som en genusneutral defaultform har formentlig været almindelig i det talte sprog i slutningen af 1800-tallet. I en beskrivelse af *-t*-formernes fremvækst (på bekostning af bl.a. former på *-n*) nævner Wiwel (1901:176) at "på dette punkt har *-t* bredt sig også foran substantiv, f. eks. ofte: "en funde*t* paraply" (eller: *-en*) (...); ligeledes "en stærkt medtage*t* bog" (sjælden *-en*)." Det er muligt at finde langt tidligere eksempler, fx "Med draget Kniv de er udlært paa Liv at stiæle" (Christian Falster: Ovidii Klage-Breve. Overs., 1738), men skal man tro de ældre grammatikker, har sådanne *-t*-former hørt til sjældenhederne før slutningen af 1800-tallet. Mere detaljerede oplysninger om *-n-* og *-t*-formernes historik kan findes i Jensen & Schack 2022.

4. Inventaret af relevante participiumsformer og adjektiver i moderne dansk

Genusbøjning af perfektum participium er mulig i den ubestemte singularisform af de participier som i bestemt form og pluralis kan have endelsen *-ne* (fx *fros**ne** ærter*, *vund**ne** kampe*). Der vil således kunne skelnes mel-

lem fx *en vunde**n** kamp* (utrum) og *et vunde**t** parti* (neutrum). Vi tager i det følgende udgangspunkt i det inventar af relevante stærke verber (og adjektiver dannet til stærke verber) som er med i Retskrivningsordbogen 2012. Ordbogen indeholder som nævnt ovenfor i alt 300 verber som kan danne sådanne participier med mulighed for en skelnen mellem *-n* og *-t*. Langt de fleste af disse verber, i alt 261, er sammensatte eller afledte/lånt fra tysk med præfiks, fx *ankomme*, *bestige*, *undertvinge*. 39 af verberne er simpleksord, fx *finde*, *skrive*, *tage*.

Udover de egentlige participiumsformer indeholder ordbogen et stort antal participiumslignende adjektiver, herunder 72 som har et tilsvarende verbum, fx *frossen* : *fryse*, *sveden* : *svide*. I nogle tilfælde svarer opslagsformen af adjektivet (dvs. utrumsformen) formelt set til en *-n*-form af participiet, fx

frossen *adj.*, -t, frosne ...
fryse *vb.*, -r, frøs, frosset *(foran fælleskønsord frossen el. frosset)* ...

Men i de fleste tilfælde fremstår *-n*-formerne som rene adjektiver fordi de ikke er opført som participiale former af et verbum, fx

forbuden *adj.*, -t, forbudne; forbuden frugt
forbyde *vb.*, -r, forbød, forbudt

Størsteparten af Retskrivningsordbogens participiumslignende adjektiver på *-n* har ikke noget verbalt modsvar, hverken i ordbogen eller i moderne dansk sprogbrug, fx *bortløben*, *fremskreden*, *fuldbåren*. Disse adjektivers neutrumsformer varierer. Nogle danner neutrum på *-t*: *et bortløbe**t** barn*. Andre danner neutrum på *-nt*: *et fuldbåre**nt** barn*. Atter andre danner neutrum på både *-nt* og *-t*: et *fremskrede**nt*** el. *fremskrede**t** tidspunkt*.

Visse forholdsvis almindeligt forekommende verber er kun repræsenteret ved et adjektiv i ordbogen, fx "**matsleben** *(el.* matslebet*) adj.*, matslebet, matslebne" (mere herom i afsnit 7.1).

5. Genusbøjning af perfektum participium ifølge retskrivningsvejledninger og -regler

Siden Dansk Retskrivningsordbog 1918 har en retskrivningsvejledning været en fast bestanddel af danske retskrivningsordbøger. I 1918-vejled-

ningen er genusbøjning af participiet obligatorisk: *en bunde**n** hest*, *et bun**det** æsel* osv. I 1955-vejledningen (Retskrivningsordbog 1955) blødes der lidt op, idet det oplyses at der er "nogen vaklen, når tillægsmåden er knyttet til et navneord." Vejledningen i formernes brug er i øvrigt meget sparsom.

Siden Retskrivningsordbogen 1986 har der været regulær valgfrihed mellem former på *-n* og *-t* foran utrumsord. Fra Retskrivningsordbogen 1986 til og med Retskrivningsordbogen 2001 nævnes valgfriheden på alfabetisk plads ved hjælp af en nøgen paragrafhenvisning, fx "**finde** *vb.*, fandt, fundet *(*funden, fundne, *jf. § 31-34)*." Den ordbogsbruger som gerne ville vide hvordan oplysningerne i parentesen skulle forstås, måtte slå efter i retskrivningsreglernes § 31-34.

I forbindelse med redigeringen af Retskrivningsordbogen 2012 blev samtlige relevante verber gennemgået. Valgfriheden mellem *-n* og *-t* blev ekspliciteret på alfabetisk plads, og formerne blev overalt forsynet med eksempler, fx

> **finde** *vb.*, -r, fandt, fundet *(foran fælleskønsord* funden *el.* fundet*)*, fundne *(jf. § 31-34)*; en funden *el.* fundet kat; et fundet dokument; fundne dyr

Ved samme lejlighed blev muligheden for genusbøjning af participiet strøget ved enkelte verber, bl.a. *komme*, hvis participiumsformer stort set aldrig bruges i attributiv stilling (i modsætning til afledte og sammensatte former som *ankommen/ankommet* og *udkommen/udkommet*).

5.1. Reglen om brugen af genusformerne – en deskriptiv regel med en implicit anbefaling

Principperne for bøjning af perfektum participium i attributiv stilling er beskrevet i retskrivningsreglernes § 31-32 (Retskrivningsordbogen 1986-2012). Af § 31 fremgår det at der kan "skelnes mellem en intetkønsform (ubøjet form) med endelsen *-et* og en fælleskønsform med endelsen *-en*" (jf. beskrivelsen ovenfor). Den "regel" som er formuleret i § 32, fremstår som en beskrivelse af sprogbrugen snarere end som en egentlig regel. Det er kun rimeligt, da det jo her drejer sig om morfosyntaks snarere end om ortografi:

> Når det gælder kønsbøjningen (…) ligger sprogbrugen dog ikke fast, men det er almindeligt at bruge intetkønsformen, også når participiet er knyttet til et substantiv af fælleskøn. I mange sådanne tilfælde kan det virke mere formelt (eller gammeldags) at bruge fælleskønsformen. (Retskrivningsordbogen 2012, § 32)

Beskrivelsen af sprogbrugen er skærpet en smule i 2012-reglens formulering: "men det er almindeligt at bruge intetkønsformen (...)." Fra 1986 til 2001 stod der i stedet "men der er en stigende tilbøjelighed til at bruge intetkønsformen (...)." Derudover er reglen uændret fra 1986 til 2012.

§ 32 kan siges at rumme en implicit anbefaling: Brug "intetkønsformen" (dvs. *-t*-formen) hvis du ikke ønsker at din tekst skal fremstå som formel eller gammeldags! Man finder samme anbefaling hos Galberg Jacobsen & Stray Jørgensen (2013:487), her blot i en mere eksplicit form: "*n*-formen virker ofte noget gammeldags, og man står sig derfor normalt ved at bruge *t*-formen."

6. Attributive *-n*-former og *-t*-former i den nutidige sprogbrug

Vi har undersøgt den nutidige brug af attributive *-n*-former og *-t*-former i et aviskorpus med ca. 1,65 mia. løbende ord (landsdækkende, regionale og lokale aviser 2004 ff.). Ud fra ususbeskrivelserne i retskrivningsreglerne og Galberg Jacobsen & Stray Jørgensen 2013 (jf. ovenfor) skulle man vente at *-n*-former i dag er sjældne i en tekstart (dvs. aviser) der ikke kan karakteriseres som "formel".

Det viser sig imidlertid at *-n*-formerne for ganske mange almindeligt forekommende participiers vedkommende står forholdsvis stærkt. I en del tilfælde forekommer *-n*-formen ligefrem hyppigere (dog sjældent signifikant hyppigere) end *-t*-formen, fx *en foretrukke**n*** snarere end *en foretrukke**t** samarbejdspartner*, *vaccine* osv., *en udeblive**n*** snarere end *en udeblive**t** effektivisering*, *menstruation* osv.

I andre tilfælde er *-n*-formerne sjældnere, fx *en forvrede**t*** snarere end *en forvrede**n** ankel*, *fod* osv., *en stjåle**t*** snarere end *en stjåle**n** elcykel*, *nummerplade* osv. For visse participiers vedkommende forekommer *-n*-formen slet ikke i vores korpus; fx finder vi kun *-t*-former i et tilfælde som *en sprunge**t** pære*, *streng* osv.

Ususbeskrivelserne i retskrivningsreglerne og Galberg Jacobsen & Stray Jørgensen 2013 holder altså kun delvis stik. Det er rigtigt at *-t*-formen *normalt* kan vikariere for *-n*-formen (jf. afsnit 6.1). Det kan derimod ikke være rigtigt at *-n*-former generelt eller overvejende opfattes som formelle eller gammeldags. Hvis det var tilfældet, ville vi næppe finde så mange *-n*-former i moderne avistekster, hvis stilleje generelt kan karakteriseres som neutralt (dvs. ikke-formelt) og nutidigt.

6.1. Principper for distributionen af *-n-* og *-t*-former i den faktiske sprogbrug

Som det er fremgået, har nogle participier typisk *-n*-form foran utrumsord, mens andre participier typisk har *-t*-form i samme kontekst. Vi har undersøgt om der med udgangspunkt i formernes distribution i den faktiske sprogbrug kan udledes nogle principper som vil kunne formidles i en kommende udgave af Retskrivningsordbogens regelafsnit. Det er selvfølgelig især interessant at se nærmere på de tilfælde som har en særlig tilbøjelighed til *-n*-former, da det jo er dem der er overraskende (med ususbeskrivelserne i afsnit 5.1 in mente).

Vi har set på følgende fire faktorer:

1. faste forbindelser (dvs. tilfælde som fx *en bunde**n/t** opgave*, *en svunde**n/t** tid*)
2. formelt/fagligt sprog (fx *bortkomme**n/t*** (formelt) vs. *forsvunde**n/t*** (neutralt)
3. participiernes stammevokaler (fx *skreve**n/t*** vs. *stjåle**n/t***)
4. telicitet (udtrykker participiet resultattilstand eller "ren" tilstand (egenskab) i den konkrete kontekst?)

De fire undersøgte mulige faktorer er valgt fordi de er nævnt i litteraturen som forhold der har eller kan have indflydelse på valget af enten *-n* eller *-t* (se nærmere om dette og om den ovenfor beskrevne undersøgelse i det hele taget i Jensen & Schack 2022).

Vi har ikke (med udgangspunkt i de fire ovennævnte faktorer) kunnet påvise nogen overordnet systematik i distributionen af *-n*-former og *-t*-former, og vi kan derfor ikke på dette grundlag fremlægge et forslag til en mere instruktiv formulering af retskrivningsreglernes § 32 (jf. afsnit

5.1). Retskrivningsordbogens redaktion bør dog nok overveje at omformulere den passus der karakteriserer *-n*-former som formelle eller gammeldags (jf. afsnit 6). Det eneste helt klare resultat af vores søgen efter distributionsprincipper er at typen "faste forbindelser" i reglen har en stærk overhyppighed af *-n*-former dér hvor det er muligt. I nogle tilfælde er *-n*-formen endda enerådende i vores korpus, fx *en svunde**n** tid* (1853 forekomster)/ *en svunde**t** tid* (0 forekomster). Forklaringen på at der her foretrækkes *-n*-former, må være at de forbindelser der giver mulighed for et valg mellem *-n* og *-t*, alle er af ældre dato. Formerne på *-n* kan derfor forklares som traderede former. Det drejer sig imidlertid om et meget beskedent antal ordforbindelser, som ikke danner grundlag for at formulere en mere generel regel. Vi hælder derfor til at forklare den overvejende del af de observerede forskelle i brugen af *-n*-former og *-t*-former som stilistisk (evt. til dels kronolektalt og/eller regionalt) betinget variation a la den variation man finder indenfor fx det ortografiske område i de tilfælde hvor der frit kan vælges mellem flere former.

7. Mulige konsekvenser for en kommende udgave af Retskrivningsordbogen

Vores undersøgelse af den faktiske sprogbrug har vist at *-n*-former stadig bruges i moderne dansk, og at de en del tilfælde endda er almindeligere end de konkurrerende *-t*-former. Vi foreslår derfor at den nuværende situation med generel valgfrihed mellem *-n*- og *-t*-former foran utrumsord fortsætter. Ved nogle participier er *-n*-former meget sjældne, og man kunne derfor overveje at opgive valgfriheden i sådanne tilfælde, fx *en sprunge**n*** el. *sprunge**t** pære* (Retskrivningsordbogen 2012) > *en sprunge**t** pære* (jf. afsnit 6). Det ville dog efter vores opfattelse kræve at man undersøgte sprogbrugen i et bredere udvalg af tekstarter. Vi har som nævnt kun undersøgt sprogbrugen i aviser, og det kan ikke udelukkes at distributionen af de to konkurrerende former er anderledes i andre tekstarter.

7.1. Forholdet mellem verber og tilsvarende adjektiver i Retskrivningsordbogen

Retskrivningsordbogen indeholder som nævnt et stort antal participiumslignende adjektiver som ikke har noget verbalt modsvar i ordbogen (jf.

afsnit 4). Mange af disse har heller ikke et verbalt modsvar i sprogbrugen, fx *hævdvunden*, *nyfalden*, *ledigbleven*. En del adjektiver modsvarer et uægte sammensat verbum, fx *hævdvunden* < *vinde hævd*.

I nogle tilfælde findes det tilsvarende verbum i sprogbrugen men ikke i ordbogen; fx har ordbogen adjektiverne *matsleben* (el. *matslebet*), *påløben*, *tilbagetrukken* (el. *tilbagetrukket*), men ikke de tilsvarende forholdsvis gængse verber, *matslibe*, *påløbe*, *tilbagetrække*. Vi foreslår at redaktionen gennemgår ordbogens udvalg af participiallignende adjektiver på *-n* og overvejer om ikke nogle af disse bør opgraderes til verber, fx *matsleben* (el. *matslebet*), adj. > *matslibe*, vb. De participiumsformer som kan dannes til verber som de tre ovennævnte, svarer til adjektivernes former, fx *matslibe* > *matsleben* foran utrumsord, *matslebet* foran neutrumsord. I andre tilfælde er der ikke fuldstændig overensstemmelse mellem verbets participiumsformer og adjektivets former. Det er emnet for det følgende afsnit.

7.2. Forholdet mellem participiers og adjektivers neutrumsformer

De relevante participier har i alle tilfælde neutrum på *-t*. I mange tilfælde har det tilsvarende adjektiv ligeledes neutrum på*-t* (jf. *matsleben* ovenfor), men i en del tilfælde har adjektivet kun neutrum på *-nt*. Vi bruger i det følgende adjektiver med *frossen* som sidsteled som typeeksempel på denne divergens mellem participium og adjektiv.

Retskrivningsordbogen indeholder 8 verber med *fryse* som sidsteled (dertil kommer simpleksverbet *fryse*). På nær *småfryse*, der typisk kun danner supinum, har alle disse verber ifølge ordbogen den almindelige mulighed for *-n* eller *-t* foran utrumsord, fx *dybfryse* (*dybfrossen* el. *dybfrosset*), *nedfryse* (*nedfrossen* el. *nedfrosset*). I de tilfælde hvor ordbogen har et tilsvarende adjektiv, ender adjektivernes neutrumsform på *-nt*. Man kan således valgfrit – og uden nogen tydelig betydningsforskel – skrive fx *et dybfrosset* (participium) *kyllingelår* og *et dybfrossent* (adjektiv) *kyllingelår*. Vi ved af erfaring at valgfriheden mellem *-t* og *-nt* i tilfælde som dette langtfra er indlysende for alle. Valgfriheden bør derfor ekspliciteres i Retskrivningsordbogen, fx ved at der fra adjektivartiklen henvises til verbumartiklen og vice versa.

7.3. Manglende adjektiver i Retskrivningsordbogen?

Typeeksemplet *frosset* (participium)/*frossen* (adjektiv) kan også bruges som illustration af mulige inkonsekvenser/vilkårligheder i Retskrivningsordbogens lemmaudvalg. Nogle af verberne med *fryse* som sidsteled har ikke noget tilsvarende adjektiv i ordbogen, og neutrumsformer på *-nt* er derfor principielt udelukket. Det gælder bl.a. *fastfryse* og *nedfryse*, som kun giver mulighed for neutrumsformer på *-t*, dvs. henholdsvis *fastfrosset* og *nedfrosset*. Formerne på *-nt* forekommer imidlertid i sprogbrugen, fx *et fastfrossent boligmarked* (Berlingske 2015), *den 4000 år gamle stump nedfrossent hår* (Fyns Amts Avis 2011). Som bekendt er ordbogsbrugere ofte meget autoritetstro, og mange vil derfor opfatte *-nt*-former som de netop nævnte som "ukorrekte". Det er imidlertid svært at give en plausibel forklaring på hvorfor sådanne former skulle være mindre "korrekte" end former som fx *dybfrossent* og *indefrossent*, som begge er mulige ifølge Retskrivningsordbogen pga. adjektivlemmaerne *dybfrossen* og *indefrossen*. Dette problem kan afhjælpes ved at redaktionen gennemgår ordbogens inventar af verber og adjektiver af typen *fryse/frossen* og supplerer med adjektiver der hvor det skønnes påkrævet.

7.4. Adjektiver med overført betydning

Når former som *dybfrossent* og *dybfrosset* står i attributiv stilling som fx i *et dybfrossent/dybfrosset kyllingelår*, kan der muligvis være en hårfin betydningsforskel (adjektivisk vs. verbal betydning), men de fleste sprogbrugere vil formentlig opfatte de to former som helt synonyme.

I nogle tilfælde er der dog en tydelig betydningsforskel på *-nt-* og *-t*-former, jf. Retskrivningsordbogen, § 35.1:

> I nogle tilfælde skelnes der betydningsmæssigt mellem et participium på *-et* og det tilsvarende adjektiv på *-en*. Participiet har da konkret betydning, mens adjektivet har overført betydning:
>
> *slebet krystal – et slebent væsen*
> *et stjålet ur – et stjålent blik*
> *et svedet øjenbryn – et svedent grin.*

Det er vores indtryk at en del (nok især yngre) sprogbrugere ikke er helt fortrolige med den ovennævnte skelnen mellem *-nt-* og *-t*-former – og måske slet ikke kender og bruger forbindelser som *et slebent væsen* og *et svedent grin*. Man ser da også at *-nt*-formen (med overført betydning) bruges i tilfælde hvor man ville forvente en *-t*-form (med konkret betydning), fx "et hifi-anlæg, som kunne tilsluttes et stjålent anlæg" (Dagbladet Roskilde 2020); "Flemming slap fra flystyrt med svedent hår, brækkede ribben og forstuvet fod" (Fyens Stiftstidende 2020). Det omvendte, dvs. *-t*-form for *-nt*-form, forekommer selvfølgelig også, fx "Hvis han lige kastede et stjålet blik på sin egen Nemkonto" (JydskeVestkysten 2021). Dette problem har dansk til fælles med norsk bokmål, som i en del tilfælde har en tilsvarende skelnen, fx *stjele – stjålet* (participium, neutrum) overfor *stjålen* (adjektiv, 'skjult, hemmelig') – *stjålent* (neutrum), jf. Bokmålsordboka. Også norske sprogbrugere kan have svært ved at holde formerne ude fra hinanden, fx "15. mai i fjor brukte hun et stjålent kort tilhørende en kvinne til å handle varer for 7232 kroner i en nettbutikk" (DA 2018).

Problemet med sammenblanding af betydningsadskillende neutrumsformer er nok vanskeligt at løse i praksis, men Retskrivningsordbogen bør i alle tilfælde eksplicitere at adjektivet har en anden betydning end det tilsvarende participium, sml. opslagene *stjålen* og *sveden* i Retskrivningsordbogen 2012:

> **stjålen** *adj.*, -t, stjålne *(skjult)*; et stjålent blik
> **sveden** *adj.*, -t, svedne; et svedent grin

I opslaget *sveden* har ordbogsbrugeren kun eksemplet *et svedent grin* at støtte sig til. Der bør, ligesom i opslaget *stjålen*, tilføjes en glosse, og der bør derudover henvises fra adjektivartiklen til verbumartiklen og vice versa. Adjektivartiklerne kunne herefter se således ud:

> **stjålen** *adj.*, -t, stjålne *(skjult)*; et stjålent blik *(jf.* stjæle*)*
> **sveden** *adj.*, -t, svedne *(listig*, *hemmelighedsfuld)*; et svedent grin *(jf.* 1. svide*)*

8. Afsluttende bemærkninger

Det er som nævnt ikke lykkedes os at finde nogen tydelig systematik i distributionen af *-nt-* og *-t*-former i moderne dansk (jf. afsnit 6.1), og vi

kan derfor ikke på grundlag af vores sprogbrugsundersøgelse fremlægge et konkret forslag til en mere instruktiv formulering af retskrivningsreglernes § 32 om bøjningen af perfektum participium foran substantiv (jf. afsnit 5.1). Vi kan derimod pege på steder i ordbogens alfabetiske del som kan gøres mere instruktive, ligesom vi foreslår at ordbogens inventar af de i denne forbindelse relevante verber og adjektiver gennemgås systematisk med henblik på at finde eventuelle lakuner i lemmabestanden (jf. afsnit 7.1-7.4).

Litteratur

Bokmålsordboka: https://ordbokene.no

Dansk Retskrivningsordbog 1918 = Viggo Saabye: *Dansk Retskrivningsordbog*. 7. udg. Med "Retskrivningsvejledning" ved Henrik Bertelsen. København: Gyldendal.

Galberg Jacobsen, Henrik & Peter Stray Jørgensen 2013. *Håndbog i Nudansk*. 6. udgave. København: Politikens Forlag.

Hansen, Erik & Lars Heltoft 2011. *Grammatik over det danske sprog*. København & Odense: Det Danske Sprog- og Litteraturselskab & Syddansk Universitetsforlag.

Jensen, Eva Skafte & Jørgen Schack 2022. Perfektum participium i attributiv stilling. Diakroni og synkroni. *Ny forskning i grammatik* 29, 84-102.

ODS = *Ordbog over det danske Sprog*. Udgivet af Det Danske Sprog- og Litteraturselskab. København: Gyldendal.

Retskrivningsordbog 1955 = *Retskrivningsordbog*. Udgivet af Dansk Sprognævn. Med "Retskrivningsvejledning" ved Erik Oxenvad og "Retskrivningsordbog" ved Jørgen Glahder. København: Gyldendal.

Retskrivningsordbogen 1986 = *Retskrivningsordbogen*. 1. udg. Redigeret og udgivet af Dansk Sprognævn. København: Gyldendal.

Retskrivningsordbogen 2012 = *Retskrivningsordbogen*. 4. udg. Redigeret og udgivet af Dansk Sprognævn. København: Alinea.

Wiwel, Hylling Georg 1901. *Synspunkter for dansk Sproglære*. København: Det nordiske Forlag.

AI-skriveassistenter og leksikografisk tekstredigering

Henrik Køhler Simonsen

Human writers using AI text generators (ATGs) seem to need help when working with an ATG and when editing texts produced by an ATG. This article analyzes and discusses how users interact with an ATG, and how they work together with an ATG to improve the text. Based on empirical data from a study with seventy test users it is demonstrated that users in fact edit in all phases of the text generation process. Data from the empirical analysis show that the editing process takes place in three phases, here referred to as the pre-editing, mid-editing, and post-editing phases. The data also seem to indicate that users need lexicographical data related to world knowledge in especially the mid-editing and post-editing phases. Based on these insights the article presents a three-phase lexicographic editing model and outlines a framework for lexicographically supported text editing.

KEYWORDS: AI text generator, text production, pre-editing, mid-editing, post-editing

1. Indledning, forskningsspørgsmål og metode

Millioner af mennesker i virksomheder og organisationer anvender i stigende grad forskellige typer teknologier til at automatisere tekstproduktionsopgaver. Især AI-baserede skriveassistenter vinder indpas (Tarp et al. 2017; Tarp 2019, 2020; Zandan 2020, Simonsen 2020a, 2020b, 2020c, 2021). Med den stigende anvendelse af disse værktøjer stiger behovet også for en dybere teoretisk forståelse af arbejdsdelingen mellem tekstproducent og den kunstige intelligens (Wilson & Daugherty 2020), samt en mere nuanceret forståelse af, hvad dette betyder for tekstproducenten og for leksikografien, (Tarp et al. 2017; Tarp 2022; Simonsen 2020a, 2020c).

En af udfordringerne ved anvendelse af AI tekstgeneratorer (ATGer) er, at tekstproducenter både skal give slip på visse opgaver, men også påtage sig helt nye opgaver. En anden af udfordringerne er, at tekstproducenter nu mere end nogensinde skal være i stand til at samarbejde med en AI og ikke mindst have sprogkompetencer nok til at kunne revidere og kvalitetssikre ATGens output, (Simonsen 2020c). En tredje udfordring er, at tekstproducenter, som anvender ATGer ikke har adgang til leksikogra-

fisk hjælp i de helt afgørende tekstredigeringsprocesser. Endelig er det helt fundamentale problem med ATGer, at de hverken trækker på leksikografisk teori og metode eller på kuraterede ordbogsdata (Simonsen 2020a, 2020b).

Alt dette betyder endvidere at fremmedsprogsundervisere nu i højere grad bør tænke nyt, udvikle nye didaktiske metoder og inddrage ATGer og leksikografiske data på en ny måde i undervisningen (Simonsen 2021, 2022; Sharples 2022).

Artiklens forskningsspørgsmål er derfor for det første at afdække og diskutere brugernes interaktion med ATGer, for det andet at udforske brugernes oplevelser med tekstredigering af den AI-genererede tekst samt for det tredje at udvikle og præsentere to modeller, som bringer leksikografien i spil i forhold til ATGer.

Artiklen er baseret på en empirisk undersøgelse, som blev gennemført i foråret 2021. I alt 115 testpersoner blev bedt om at teste en specifik ATG. Undersøgelsen er baseret på et sample på 70 personer, dvs. N=70, hvoraf 42 testpersoner var kommunikationseksperter og 28 testpersoner var kommunikationsstuderende. De 70 testpersoner løste først en opgave ved hjælp af ATGen Sassbook, hvor de fik til opgave at producere en blogtekst om e-bøger, og derefter deltog de i en survey med fem kvantitative spørgsmål og tre kvalitative spørgsmål.

Dette genererede kvantitative data om brugernes oplevelser målt på en skala fra 1 til 5 samt 210 kvalitative udsagn. De kvantitative data blev udsat for en ikke-parret, dobbeltsidet T-test, som viste, at de indhentede data var statistisk signifikante. De kvalitative data blev læst ind i NVivo, og der blev gennemført en CAQDAS-understøttet tematisk analyse ved hjælp af de tre koder "AI text generator", "Cooperation" og "Editing", som alle tre fokuserer på artiklens problemstilling.

Ved hjælp af NVivo og de tre koder var det muligt at kategorisere udsagnene efter henholdsvis tema, frekvens og kontekst, og de NVivo-genererede analyser viste temaerne i kontekst, hvilket gjorde det muligt at diskutere artiklens forskningsspørgsmål.

De empiriske data har tilvejebragt en række indsigter, som har været instrumentale i besvarelsen af forskningsspørgsmålene og i udviklingen af de leksikografiske overvejelser.

2. AI-skriveassistenter og leksikografi

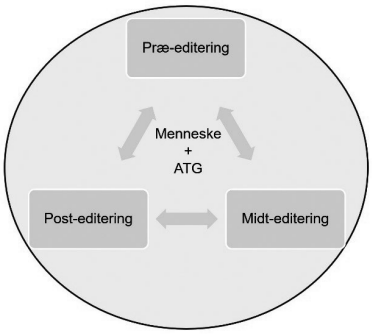
En af leksikografiens styrker har altid været at udvikle værktøjer, som har bestemte formål og hjælper brugerne med f.eks. at lære og tilegne sig viden (kognitiv funktion) samt skrive, tale og kommunikere (kommunikativ funktion), jf. f.eks. Wiegand (1997:194), som taler om, at "Jeder Gebrauchsgegenstand hat mindestens einen genuinen Zweck […]". Det har i mange år været et af leksikografiens bærende principper at hjælpe brugeren med at opnå specifikke handlingsmål, og leksikografien har da også siden midt-80erne været særligt optaget af dette. Arbejdet med sprog, retning og korrektur har også i lang tid været et af leksikografiens interesseområder, se f.eks. Tarp (2004), som foreslår korrektur og retning som deciderede leksikografiske funktioner, Tarp (2008), som præsenterer en teori for "learner lexicography", samt Tarp (2019), som på basis af arbejdet med skriveassistenten Writeassistant, foreslår nytænkning af leksikografien i kølvandet på de disruptive skriveassistenter. Et andet bidrag om skriveassistenter er Fuertes-Olivera & Tarp (2020), som foreslår, at leksikografien skal udgøre en vigtig teoretisk byggeklods i de fremadstormende skriveassistenter. Endelig har sammensmeltningen mellem leksikografi og de AI-baserede skriveassistenter været temaet i flere bidrag (Simonsen 2020a, 2020b, 2020c, 2021, 2022), som har argumenteret for, at leksikografisk metode og teori, samt ikke mindst kuraterede leksikografiske data, skal være de bærende elementer i moderne, brugerorienterede leksikografiske ATGer.

Automatisk tekstgenerering er her nemlig allerede, og millioner af både studerende og medarbejdere bruger ATGer til at skrive tekster med (Sharples 2022). Analogt med Sharples (2022), som argumenterer for, at undervisere nu er tvunget til at gentænke den måde de underviser og evaluerer på, argumenterer Simonsen (2021:240) at "… working with an AI demands a lot from the students as they will have to spend much more time on high-cognitive processes such as pre-editing, mid-editing and post-editing texts".

Tidlige undersøgelser af ATGer (Simonsen 2020a, 2020b) synes ligeledes allerede for tre år siden at have afsløret, at der er særlig brug for nye teorier og metoder til at hjælpe brugeren med at forbedre det indhold som ATGen leverer. En ATG har nemlig ikke omverdensviden og en ATG kan stadig ikke skabe kausale sammenhænge, inddrage omverdensviden i

genereringen af tekst eller opbygge argumentationsrækker.

Simonsen (2021:240) peger konkret på "Up until now, we have mainly focused on the student's ability to write correct and coherent texts in text production classes, but we will have to change that focus in our future curricula to prepare students for an AI-intensive world." Simonsen (2021:240) præsenterer endvidere den trefasede redigeringsmodel og argumenterer, at tekstredigering i forbindelse med ATGer synes at foregå i tre faser som vist i figur 1.



FIGUR 1. Trefaset redigeringsmodel (Simonsen 2021:240).

Denne trefasede opdeling anvendes som udgangspunkt for analysen af data fra de 70 testpersoner og i den følgende diskussion vil begreberne omverdens-, redigerings- og sprogkompetencer blive anvendt om henholdsvis brugerens evne til at inddrage omverdensviden i redigeringsfaserne, brugerens evne til at foretage kognitiv redigering og endelig brugerens evne til at anvende sproglige redigeringsevner.

Ifølge Simonsen (2021) involverer præ-editering en række vigtige overvejelser fra brugerens side. Brugeren skal f.eks. gøre sig overvejelser om tekstens kommunikative formål, sprogretning, tone of voice samt genre. Derudover skal brugeren i præ-editeringsfasen nøje kunne udvælge den tekst eller de nøgleord, som ATGen skal anvende for at komme i gang. Brugeren skal således anvende både omverdens-, redigerings-, og sprogkompetencer for at starte ATGen.

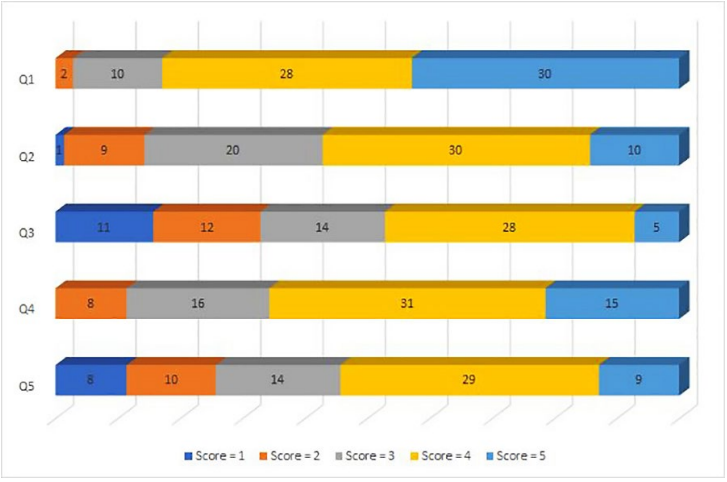
Midt-editering betyder ifølge Simonsen (2021), at brugeren i særlig grad skal udvælge foreslåede løsninger samt slette uønskede forslag og gennemføre rettelser i teksten. Det kan også indebære, at brugeren skal tilføje nye nøgleord for at føre den automatiske tekstproduktion tilbage

på sporet og dermed styre retningen for og indholdet i den endelige tekst. Også her skal brugeren anvende omverdens-, redigerings-, og sprogkompetencer for at styre ATGen.

Endelig argumenterer Simonsen (2021), at post-editeringsfasen er særligt opmærksomhedskrævende og at brugeren i særlig grad skal være i stand til at anvende sine sprogkompetencer til at rette og forbedre den tekst, som ATGen har produceret. Sprogkompetencer i og viden om sprogets grammatiske, syntaktiske og stilistiske træk er helt afgørende, men også generelle redigerings- og omverdenskompetencer er nødvendige for at kunne post-editere.

3. Analyse og diskussion

De 70 testpersoner testede først ATGen. Testpersonerne klikkede på et link til et dokument med instruktioner og en færdig tekst, som de skulle bruge. Forsøget var uobserveret, men det kan ses på besvarelserne, at alle testpersoner rent faktisk har testet ATGen. Testpersonerne fik til opgave at skrive en kort blogtekst, hvis kommunikative formål var at informere og overbevise om fordelene ved e-bøger. Kanalen for teksten var en fiktiv virksomheds LinkedIn-profil. Da testpersonerne havde produceret en tekst om e-bøger, svarede de på fem kvantitative spørgsmål. Figur 2 viser de 70 testpersoners svar på de fem kvantitative spørgsmål på en skala fra 1 til 5 i et farvemarkeret, stablet søjlediagram.



FIGUR 2. Testpersonernes svar på de fem spørgsmål Q1-Q5.

Da testpersonerne i Q1 blev spurgt om, hvor brugervenligt de oplevede ATGen svarede de fleste, at de var enten "tilfreds" eller "meget tilfreds", som det fremgår af henholdsvis den gule og blå farve i den øverste liggende søjle. Da de i Q2 blev spurgt om kvaliteten af den genererede tekst var svaret en del mere negativt, som vist i den anden liggende søjle set fra oven. En af grundene til denne mere forsigtige bedømmelse er formentlig den manglende transparens i en ATG, som flere af testpersonerne bemærkede i den kvalitative del. Brugerne kan ikke se, hvor teksten kommer fra, og om den er valid. Denne indsigt er vigtig for udviklingen af en leksikografisk redigeringsinterface, som det vil fremgå senere.

Da testpersonerne i Q3 blev spurgt om, hvorvidt de ville bruge en ATG i deres nuværende stilling eller studier var billedet noget mere uklart, og selvom hele 28 af de adspurgte har scoret 4 som det fremgår af den gule farve, så har mange også svaret, at de ikke kan se sig selv anvende en ATG. De kvalitative data understøtter dette, og igen skyldes det primært, at testpersonerne ikke kan se, hvor teksten kommer fra, og om kilden er troværdig. Faktisk spørger en af testpersonerne til hvilken kilde, der er anvendt. Da testpersonerne i Q4 blev spurgt om, hvorvidt de oplevede ATGen som en værdiskabende teknologi var svaret overvejende positivt, som det fremgår af den fjerde liggende søjle set fra oven. Da testpersonerne i Q5 blev spurgt om de kunne se anvendelsesmulighederne på jobbet var svaret, at over halvdelen af testpersonerne ville bruge ATGen i professionelt.

Disse indsigter er instrumentale i forståelsen af ATGer og dermed også et helt centralt element i diskussionen af, hvordan leksikografien kan forbedre ATGer. Uden viden om ATGerne er det vanskeligt at udvikle nye leksikografiske løsninger som kan hjælpe brugerne i de tre redigeringsfaser.

Efter at have testet ATGen svarede testpersonerne på tre udvalgte kvalitative spørgsmål "What tasks in the text production process did the AI Writer solve?", "What tasks in the text production process did you as a human solve?" og "What do you need to be able to work with an AI writer?" og der blev gennemført en CAQDAS-understøttet tematisk analyse ved hjælp af NVivo. Som allerede anført ovenfor, var den tematiske analyse centreret omkring de tre koder "AI text generator", "Cooperation" og "Editing". De i alt 210 udsagn blev uploaded og analyseret ved hjælp af NVivo og følgende udvalgte udsagn i tabel 1 beskriver, hvordan testpersonerne oplevede arbejdsprocessen i præ-editeringsfasen:

TABEL 1. Udsagn om præ-editeringsfasen

I created the foundation and keywords for the text
It is a quick and easy way of going from keywords to actual written content
Picking the subject matter and some descriptive words
Provide the AI writer with quality input in order to get quality results

De kvantitative data og de udvalgte kvalitative udsagn peger på, at langt størstedelen af testpersonerne var forbløffede over, hvordan ATGen kom fra enkelte nøgleord til en færdig tekst. Data synes også at pege på, at denne opgave er særlig vigtig, fordi nøgleordenes kvalitet har stor betydning for, om ATGen kan generere tilfredsstillende output. Også her er brugerens omverdenskompetencer vigtige, dvs. det er vigtigt at brugeren kan navigere i og finde de rigtige tekster, hvorfra relevante og præcise nøgleord kan uddrages og anvendes i ATGen. Dette peger direkte på det foreslåede leksikografiske redigeringskoncept.

Følgende udsagn i tabel 2 beskriver, hvordan testpersonerne oplevede forløbet under selve tekstgenereringen og den redigering, der finder sted mens ATGen genererer tekst:

TABEL 2. Udsagn om midt-editeringsfasen

I changed the format to make the text more coherent
I did have to edit a few things such as deleting a link, which was in the middle of a sentence, and I edited some half-finished sentences
I removed too many empty spaces and wrong commas

Både de kvantitative data og de kvalitative udsagn synes at vise, at det er nødvendigt at foretage bestemte valg mens ATGen genererer selve teksten. Som det fremgår ovenfor, er det dog mindre redigeringsopgaver, der skal gennemføres, men dette arbejde er vigtigt for at holde ATGen "på sporet". I denne fase er der især brug for brugerens redigeringskompetencer og sprogkompetencer og også her føder indsigterne direkte ind i det foreslåede redigeringskoncept i afsnit 4.

Endelig udtalte testpersonerne sig også om, hvordan de arbejdede med den AI-genererede tekst:

TABEL 3. Udsagn om post-editeringsfasen

The AI writer provided suggestions and automatically typed the text, so I only had to post-edit the output.
I provided some "feelings" and "compassion" via creative writing
I proofread the text and added context to the text and ensured the red thread throughout the text

Data synes at vise, at testpersonerne er tilfredse med ATGen, men at de ikke er tilfredse med kvaliteten af den AI-genererede tekst. Dette er særligt interessant for det leksikografiske redigeringskoncept. Data pegede også på, at testpersonerne redigerer i processen, men at de i ingen af faserne har adgang til den type hjælp, som de har mest brug for – nemlig leksikografiske data. Som udsagnene viser, har brugerne brug for både omverdens-, redigerings-, og sprogkompetencer.

Det handler både om at inddrage viden om omverden samt om den kognitive og kommunikative redigering af teksten. Men hjælp til de forskellige former for redigering får brugeren ikke. Der synes således både at være brug for et leksikografisk korpus (kuraterede leksikografiske data som en del af ATGens hjerne) og et indbygget leksikografisk interface ovenpå ATGen for at hjælpe brugeren med de helt afgørende redigeringsopgaver både før, under og efter. ATGen har nemlig ikke nogen omverdensviden og har særlig brug for denne menneskelige hjælp for at den genererede tekst kan forbedres. Der er brug for leksikografiske data som redigeringshjælp.

4. Leksikografiske data som redigeringshjælp i AI-skriveassistenter

På basis af indsigterne fra analysen og diskussionen, og med inspiration fra (Tarp 2004, 2008; Leroyer & Simonsen 2019; Simonsen 2008) præsenteres et koncept, som kan tilfredsstille brugernes behov for redigeringshjælp. Udgangspunktet for konceptet er, at dokumenter ses som leksikografiske ressourcer, jf. Simonsen (2008), som præsenterer en "lexicographic document template model", altså et forslag om, at brugeren ved søgning efter et lemma også får adgang til en nøje udvalgt og kurateret tekst, hvori opslagsordet optræder.

Ifølge Simonsen (2008:1066-1067) indebærer dette, at brugeren i forbindelse med opslagsgerningen får adgang til information om tekstens kommunikative formål (som hjælper brugeren med at imitere stil og tone), information om den typiske trækstruktur eller genretræk for den pågældende teksttype samt information om nogle af de typiske retoriske strategier i den pågældende teksttype.

En lignende tilgang foreslås anvendt i udviklingen af et leksikografisk redigeringsinterface. Dette er dog kun en midlertidig løsning, indtil de leksikografiske miljøer i samarbejde med relevante partnere udvikler ægte leksikografiske ATGer, hvor de leksikografiske data ikke kun er en del af ATGens hjerne men også indeholder vigtig omverdensviden. De tre typer af data til realisering af henholdsvis præ-editering, midt-editering, og post-editering og de tilhørende adgange til tekster etc. foreslås implementeret i ATGer i relevante dele af skærmbilledet. Selvom ATGer bliver bedre, er redigering vigtigere end nogensinde før, og tabellerne herunder viser både behovet (lysegrå baggrund) og forslag til udvalgte leksikografiske data og tekster i de tre redigeringsfaser (hvid baggrund).

Den første gruppe af indsigter fra undersøgelsen peger på, at det er vigtigt, at brugeren får hjælp til præ-editeringsfasen som vist i tabel 4.

TABEL 4. Leksikografiske data til realisering af præ-editeringsfunktion

Brug for hjælp til at finde kuraterede og kvalitetssikrede nøgleord og sætninger
Adgang til kuraterede og kvalitetssikrede emne- og situationsspecifikke tekster og anvende kognitive og kommunikative omverdens- og sprogkompetencer og dermed ekstrahere relevante nøgleord og sætninger.
Brug for hjælp til at vælge rigtig genre, tone og stil
Adgang til kuraterede og kvalitetssikrede genre-, tone-, og stilspecifikke tekster og anvende kognitive og kommunikative omverdens- og sprogkompetencer og dermed ekstrahere relevante nøgleord og sætninger.

Den anden gruppe af indsigter peger på, at det også er vigtigt, at brugeren får hjælp til holde ATGen på sporet under selve tekstgenereringen som vist i tabel 5.

TABEL 5. Leksikografiske data til realisering af midt-editeringsfunktion

Brug for hjælp til at sortere i forslag, som ATGen tilbyder
Adgang til kuraterede og kvalitetssikrede emne- og situationsspecifikke tekster og anvende kommunikative og kognitive omverdens-, redigerings-, og sprogkompetencer og dermed udvælge, kopiere, slette og redigere.
Brug for hjælp til kommunikativ korrektur
Adgang til leksikografiske data (ortografi, grammatik, syntaks, etc.) af især kommunikativ karakter og anvende kommunikative omverdens-, redigerings-, og sprogkompetencer og dermed udvælge, kopiere, slette og redigere.
Brug for hjælp til kognitiv korrektur
Adgang til leksikografiske data (eksempler, definitioner, etc.) af især kognitiv karakter og anvende kognitive omverdens-, redigerings-, og sprogkompetencer og dermed udvælge, kopiere, slette og redigere.

Den sidste gruppe af indsigter viser, at der er brug for leksikografiske data i post-editeringsfasen, hvor brugeren behøver hjælp til både kommunikativ og kognitiv korrektur som foreslået i tabel 6.

TABEL 6. Leksikografiske data til realisering af post-editeringsfunktion

Brug for hjælp til kommunikativ korrektur
Adgang til leksikografiske data (ortografi, grammatik, syntaks, etc.) af især kommunikativ karakter og anvende kommunikative omverdens-, redigerings-, og sprogkompetencer og dermed udvælge, slette og redigere.
Brug for hjælp til kognitiv korrektur
Adgang til leksikografiske data (eksempler, definitioner, etc.) af især kognitiv karakter og anvende kognitive omverdens-, redigerings-, og sprogkompetencer og dermed udvælge, slette og redigere.

5. Konklusioner og perspektiver

Undersøgelsen påviste hvordan de 70 testpersoner oplevede at arbejde med en ATG, og det argumenteres, at indsigterne har bidraget til vores forståelse af arbejdsdelingen mellem en menneskelig tekstproducent og en ATG. Undersøgelsen viste også, hvordan brugerne løste forskellige redi-

geringsopgaver både før, under og efter og undersøgelsen synes således at bekræfte hypotesen om, at brugere redigerer i mindst tre faser.

Undersøgelsen af hvordan brugere anvender ATGer er en forudsætning for at forstå, hvor leksikografien kan bidrage. På basis af de identificerede indsigter var det muligt at udvikle en trefaset redigeringsmodel samt et leksikografisk redigeringskoncept, som i overensstemmelse med den enkelte redigeringssituation præsenterer nøje udvalgte leksikografiske data i de tre faser før-under-efter.

ATGer er en realitet og en alvorlig konkurrent til tekstproduktionsordbøger. Men der synes stadig mere end nogensinde at være brug for leksikografisk teori, leksikografiske data og leksikografisk metode – også i ATGer. Leksikografien skal mere på banen og være med til at forbedre ATGer og denne artikel er et bidrag til denne vigtige udvikling.

Litteratur

Fuertes-Olivera, Pedro Antonio & Sven Tarp 2020. A Window to the Future: Proposal for a Lexicography-assisted Writing Assistant. *Lexicographica – International Annual for Lexicography*, 36(1), 257-286.

Leroyer, Patrick & Henrik Køhler Simonsen 2019. Google Translate: trussel eller redning for oversættelsesordbøger? *LexicoNordica* 26, 95-115.

Sharples, Mike 2022. New AI tools that can write student essays require educators to rethink teaching and assessment. <https://blogs.lse.ac.uk/impactofsocialsciences/2022/05/17/new-ai-tools-that-can-write-student-essays-require-educators-to-rethink-teaching-and-assessment/>. Hentet juli 2022.

Simonsen, Henrik Køhler 2008. Lexicographic Document Templates: Text Genre Conventions in Lexicography. I: Bernal, Elisenda & Janet DeCesaris (red.), *Proceedings of the XIII EURALEX International Congress*. 15-19 July 2008. Barcelona: Universitat Pompeu Fabra, 1065-1072.

Simonsen, Henrik Køhler 2020a. Augmented Writing: nye muligheder og nye teorier. I: Sandström, Caroline, Ulla-Maija Forsberg, Charlotta af Hällström-Reijonen, Maria Lehtonen & Klaas Ruppel (red.), *Nordiska studier i lexikografi* 15. Helsingfors: Nordisk förening för lexikografi, 307-315.

Simonsen, Henrik Køhler 2020b. Når Augmented Writing og leksikografi går hånd i hånd. *LEDA-nyt* nr. 69, 3-13.

Simonsen, Henrik Køhler 2020c. Augmented Writing Needs Lexicography. I: Gavriilidou, Zoe, Maria Mitsiaki & Anna Fliatouras (red.), *Proceedings of XIX EURALEX Congress: Lexicography for Inclusion*, Vol. I. Alexandroupolis: Democritus University of Thrace, 509-514.

Simonsen, Henrik Køhler 2021. AI Writers in Language Learning. I: Chang, Maiga, Nian-Shing Chen, Demetrios G. Sampson & Ahmed Tlili (red.), *Proceedings IEEE 21st International Conference on Advanced Learning Technologies*, Los Alamitos, CA: IEEE, 238-240.

Simonsen, Henrik Køhler 2022. AI Text Generators and Text Producers. I: Chang, Maiga, Nian-Shing Chen, Demetrios G. Sampson & Ahmed Tlili (red.), *Proceedings IEEE 22nd International Conference on Advanced Learning Technologies,* Los Alamitos, CA: IEEE, 218-220.

Tarp, Sven 2004. Korrektur og retning som leksikografiske funktioner. *Hermes – Journal of Linguistics* 33, 117-147.

Tarp, Sven 2008. *Lexicography in the Borderland between Knowledge and Non-Knowledge: General Lexicographic Theory with Particular Focus on Learner's Lexicography*. Tübingen: Max Niemeyer.

Tarp, Sven 2019. Connecting the Dots: Tradition and Disruption in Lexicography. *Lexikos* 29, 224-249.

Tarp, Sven 2020. Integrated Writing Assistants and their Possible Consequences for Foreign-Language Writing and Learning. I: Ana Bocanegra Valle (red.), *Applied linguistics and knowledge transfer: employability, internationalization and social challenges*. (Linguistic Insights: Studies in Language and Communication 268.) Bern: Peter Lang, 53-76.

Tarp, Sven 2022. Turning Bilingual Lexicography Upside Down: Improving Quality and Productivity with New Methods and Technology. *Lexikos* 32, 66-87.

Tarp, Sven, Kasper Fisker & Preben Sepstrup 2017. L2 writing assistants and context-aware dictionaries: New challenges to lexicography. *Lexikos* 27, 494-521.

Wiegand, Herbert Ernst 1997. Über die gesellschaftliche Verantwortung der wissenschaftlichen Lexikographie. *Hermes – Journal of Linguistics* 18, 177-202.

Wilson, James & Paul Daugherty 2020. Collaborative Intelligence: Humans and AI are Joining Forces. <https://hbr.org/2018/07/collaborative-intelligence-humans-and-ai-are-joining-forces>. Hentet juli 2022.

Zandan, Noah 2020. The Future of Human Communication: How Artificial Intelligence Will Transform the Way We Communicate <https://www.quantifiedcommunications.com/blog/artificial-intelligence-in-communication>. Hentet juli 2022.

Kva gjer ordbøker når rettskriving, ordklassar og til og med kommunegrenser endrar seg?

Klara Sjo & Gyri Smørdal Losnegaard

This article describes how the dictionary updating projects at the University of Bergen deal with changes in 1) spelling, 2) grammar and 3) geographical-administrative borders, in view of the type of dictionary, publication format and audience. NO-AH revises the documentation dictionary *Norsk Ordbok*, and Revisjonsprosjektet revises *Nynorskordboka* and *Bokmålsordboka*, the standard spelling dictionaries of Norwegian.

An important factor in addressing changes and meeting user expectations is to exploit the possibilities in online publishing. By connecting the dictionaries to digital resources like ordbanken and Metaordboka, changes in orthography and part of speech updates automatically and quickly. *Norsk Ordbok* uses geographical references, and we outline different ways of presenting geographic data to deal with changing geographical borders and administrative names. A solid and flexible "did you mean"-feature helps find the correct word form when searching in *Nynorskordboka* and *Bokmålsordboka*. A similar solution could be used to find the official forms when users search with outdated forms in these dictionaries, and to find dialect forms in *Norsk Ordbok*.

NØKKELORD: nettordbøker, revisjon, rettskriving, ordklasse, geografi

1. Innleiing

I 2016 overtok Universitetet i Bergen (UiB) standardordbøkene *Bokmålsordboka* og *Nynorskordboka* (ordbøkene.no) og dokumentasjonsordboka *Norsk Ordbok* (NO) frå Universitetet i Oslo. Dei tre ordbøkene blir i dag oppdaterte i to revisjonsprosjekt: *Bokmålsordboka* og *Nynorskordboka* i Revisjonsprosjektet (2018–2024) og *Norsk Ordbok* i prosjektet NO-AH (2019–2029). Opphaveleg var dette trykte ordbøker med form og innhald tilpassa papirformatet som etter kvart fekk nettutgåver: standardordbøkene i 1994 og NO i 2014. Etter overføring til UiB er ordbøkene heildigitale, og dei blir no utelukkande skrivne og tilrettelagde for visning på nett.

Med nettsidene som einaste publiseringsplattform er det ei sentral oppgåve i begge prosjekt å skape fullverdige digitale ordbokressursar. Men revisjonen byr på problemstillingar knytt til endringar i rammeverk som ordbøkene bygger på: Korleis og i kva grad skal retningsliner og praksis i redigeringa av ordbøkene justerast når forhold har endra seg sidan dei blei skrivne? Her spelar både publiseringsformat, typen ordbok og målgruppa til den respektive ordboka inn.

I artikkelen tar vi føre oss endringar i 1) offisiell rettskriving, 2) grammatiske kategoriar og 3) geografisk-administrative inndelingar. Vi viser korleis strukturelle endringar på desse ulike områda påverkar dei to ordboksprosjekta, og korleis dei blir handterte i dei ulike ordbøkene.

2. Om prosjekta og ordbøkene

Bokmålsordboka og *Nynorskordboka* blir no gjennomgått fullstendig, for første gang sidan dei blei skrivne på byrjinga av 1980-talet. Standardordbøkene er normgjevande og brukte som referanse for dei som skal skrive normert norsk i skule, lærebøker og offentleg forvalting. UiB og Språkrådet eig standardordbøkene i fellesskap. UiB har det redaksjonelle ansvaret, og Språkrådet som normgjevande organ står på eigarsida både i standardordbøkene og Norsk ordbank (ordbanken), databasen som ordbøkene hentar rettskrivingsopplysningar frå. Språksamlingane ved UiB driftar ordbøkene og tilknytte databasar.

Standardordbøkene dekker sentrale delar av ordforrådet. Artiklane viser bøying, ein relativt kortfatta definisjon, eksempel på bruk og eventuelt ein kort etymologi og uttale. Hovudmålsetjinga for prosjektet er å gjere utvalet av oppslagsord i *Bokmålsordboka* og *Nynorskordboka* likare, sørge for at innhaldet er i tråd med dagens språkbruk, og ta inn nye oppslagsord og tydingar. Etter revisjonen vil bøkene ha eit omfang på om lag 100 000 ord kvar.

I 2021 og 2022 hadde vi to brukarundersøkingar på nettsidene (sjå Rauset 2022), med rundt 7000 respondentar på kvar. Den primære målgruppa for ordbøkene.no er av UiB definert som studentar og undervisarar i heile utdanningsløpet (Rauset 2022), men svara i undersøkinga tyder på ei meir samansett brukargruppe. I respondentgruppa utgjer saksbehandlarar, dei som skriv som ein del av jobben og omsetjarar 34 %, elevar og studentar 21 % og undervisarar 16 %. Vidare definerer 22 % seg som

språkinteresserte og 7 % anna. Ikkje minst seier heile 20 % at dei har eit anna førstespråk enn norsk, ei ny gruppe det er viktig å legge til rette for.

Prosjektet Norsk Ordbok a–h (NO-AH) reviderer alfabetstrekket a–h i NO, som er ei dokumentasjonsordbok over det nynorske skriftspråket og dei norske dialektane. Ordboka bygger på ei omfattande kjeldesamling av både litterære kjelder og målføredokumentasjon. UiB eig ordboka og dei fysiske og digitale samlingane, som er forvalta av Språksamlingane. NO er omfangsrik og informasjonstung. Ordboka har meir enn 330 000 artiklar og inneheld mange typar opplysingar: om kjelder, uttale, geografisk plassering og utbreiing av dialektord. Målgruppa er forskarar innan humaniora, lærarar og lekfolk med interesse for språk og språktradisjon (Vikør 2018:10). Bruk av ordboka krev bakgrunnskunnskap, og brukarane vil vere relativt avanserte samanlikna med brukarane av ei standardordbok. Det digitale rammeverket til NO er under utvikling og dei tekniske løysingane skal bygge på rammeverket som er utvikla ved UiB for ordbøkene.no.

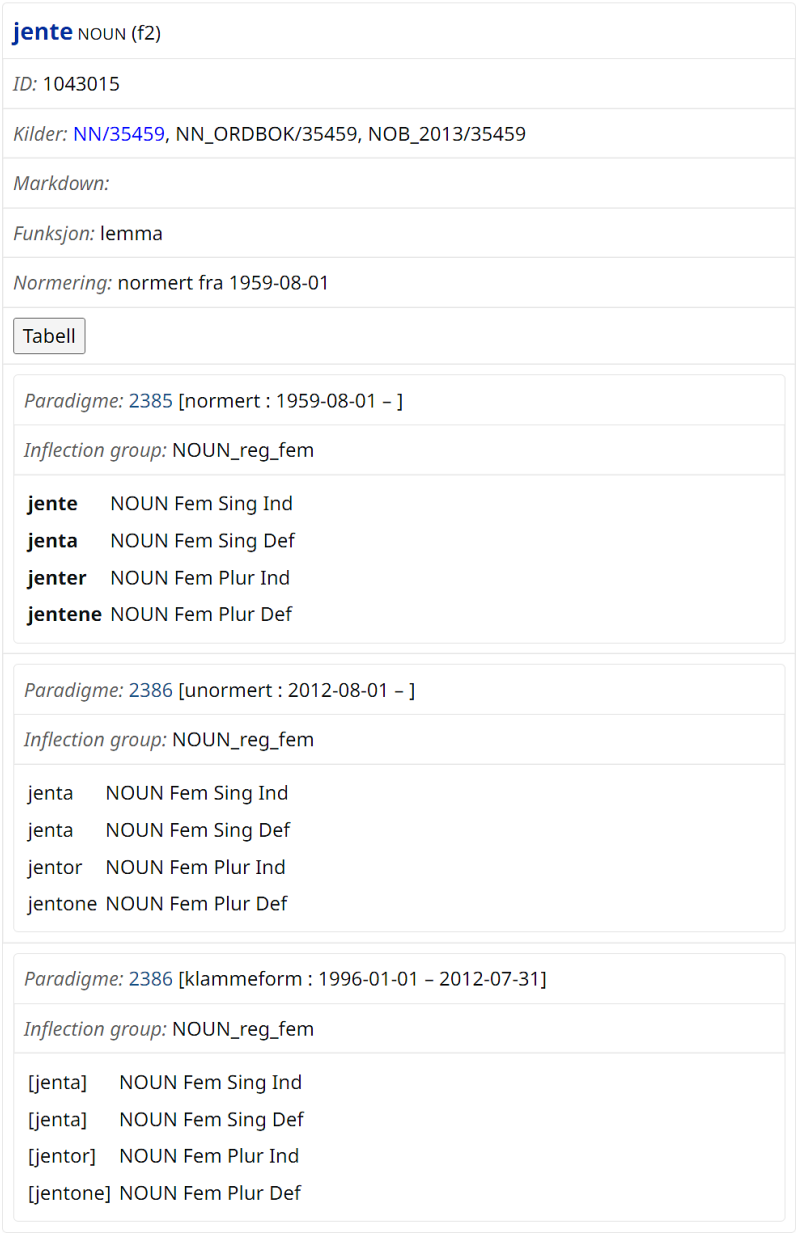
3. Offisiell rettskriving

Norsk er kjenneteikna av ei norm med stor valfridom og relativt hyppige språkreformer. Dei siste store språkreformene var i 2005 for bokmål og 2012 for nynorsk, men det kjem mindre endringar med jamne mellomrom, både for rettskriving og bøying.

Bokmålsordboka og *Nynorskordboka* skal som normgjevande ordbøker vise dei gjeldande skrivemåtane og bøyingsmønstera for alle ord. Språkrådet fastset skrivemåte og bøyingsmønster for ord, og desse blir lagde inn i ordbanken, som ordbøkene.no er knytte til. Ordbanken blei opphaveleg utvikla ved Universitetet i Oslo og samlar rettskrivingsinformasjon for norsk språk (Grønvik & Ore 2014). Ordbanken er to databasar – ein for nynorsk og ein for bokmål – av grunnord og bøyingsmønster og kopling mellom dei. Ordbanken lenker kvar grunnform til eitt eller fleire bøyingsmønster. *Nynorskordboka* blei kopla til ordbanken i 2012 og *Bokmålsordboka* i 2015.

Koplinga mellom ordbanken og ordbøkene.no gjer at alle endringar i rettskriving og bøying blir automatisk oppdaterte i standardordbøkene. Oppdatering og vedlikehald av ordbanken føregår hos UiB som saman med Språkrådet har det overordna ansvaret for utviklinga av ressursen.

Ordbanken viser òg historikken til eit ord, altså skrivemåtar og former som er gått ut av norma (sjå figur 1), men berre dei normerte formene blir viste i ordbøkene.



jente NOUN (f2)

ID: 1043015

Kilder: NN/35459, NN_ORDBOK/35459, NOB_2013/35459

Markdown:

Funksjon: lemma

Normering: normert fra 1959-08-01

Tabell

Paradigme: 2385 [normert : 1959-08-01 –]

Inflection group: NOUN_reg_fem

jente	NOUN Fem Sing Ind
jenta	NOUN Fem Sing Def
jenter	NOUN Fem Plur Ind
jentene	NOUN Fem Plur Def

Paradigme: 2386 [unormert : 2012-08-01 –]

Inflection group: NOUN_reg_fem

jenta	NOUN Fem Sing Ind
jenta	NOUN Fem Sing Def
jentor	NOUN Fem Plur Ind
jentone	NOUN Fem Plur Def

Paradigme: 2386 [klammeform : 1996-01-01 – 2012-07-31]

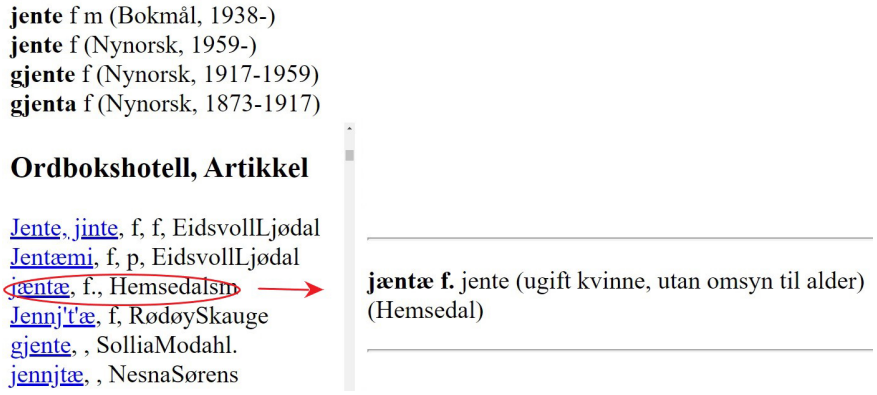
Inflection group: NOUN_reg_fem

[jenta]	NOUN Fem Sing Ind
[jenta]	NOUN Fem Sing Def
[jentor]	NOUN Fem Plur Ind
[jentone]	NOUN Fem Plur Def

FIGUR 1. Oppslaget *jente* i ordbanken, som viser normerte og tidlegare normerte former samt klammeform (alternativ normert form).

Redigeringa av *Norsk Ordbok* starta på 1930-talet, og ordboka brukar rettskrivingsnormalen frå 1938. Mens dagens rettskriving f.eks. tillet både a- og e-infinitiv av verb (*hoppa* og *hoppe*), vil ein ikkje få treff i NO om ein søker med e-infinitiv sidan slike former ikkje var del av 1938-normalen. I Språklova frå 2021 blir det stilt krav om at NO skal gje opplysingar om gjeldande offisiell rettskriving i artikkelhovudet.

Norsk Ordbok hentar opplysingar om rettskriving frå Metaordboka (MO), ein leksikalsk database utvikla for ordboka og som fungerer både som rettskrivingsressurs og redigeringsverktøy (Ore 2000, Grønvik & Ore 2018). Mesteparten av kjeldene til *Norsk Ordbok* er digitaliserte, og Metaordboka fungerer som ein felles søkeinngang til dei. Alle registrerte variantar av eit ord, skriftleg eller munnleg, er samla i ein MO-artikkel, og kvar artikkel i *Norsk Ordbok* er oppretta med utgangspunkt i ein MO-artikkel. Figur 2 viser eit utsnitt av artikkelen for substantivet *jente* slik han ser ut i MO. Heilt oppe til venstre ser vi opplysingar om tidlegare og gjeldande rettskriving. Skriftforma *gjente* med *g* var normert form i 1938, men gjekk ut av rettskrivinga i 1959, mens *jente* med *j* har vore normert sidan 1938 – og er det framleis. MO blir no oppdatert med 2012-rettskrivinga for nynorsk.



FIGUR 2. Utsnitt av MO-artikkelen for substantivet *jente*.

Til venstre i figur 2, under normeringsinformasjonen, ser vi variantar av ordet *jente* i ulike kjelder. Ordbokshotellet er ein kjeldedatabase som inneheld både ordbøker, ordlister og dialektsamlingar. Under denne overskrifta ser vi bl.a. at uttalevarianten *jæntæ* er belagt i ei dialektordsamling frå Hemsedal. Den aktuelle artikkelen frå ordsamlinga er vist til høgre.

Oppslagsformer i *Norsk Ordbok* er hovudformene frå 1938, og ein søker i ordboka med denne rettskrivinga. Det er ein del skilnader mellom dagens rettskriving og 1938-rettskrivinga, og brukarane er ikkje nødvendigvis kjende med desse, noko som kan vere ei utfordring. Kravet om at ordboka skal vise gjeldande offisiell rettskriving inneber at Metaordboka må oppdaterast med informasjon om kva som er gjeldande normert form og eventuelt bøyingsmønster. Tidlegare normerte og unormerte former må merkast. Dei siste åra er det gjort eit større arbeid med å oppdatere Metaordboka med oppslagsform etter gjeldande norm, jamfør artikkelen for *jente*. Ein viktig del av moderniseringa av NO blir å kople saman Metaordboka og ordbanken slik at ordboka kan vise oppdatert rettskriving og bøying.

4. Grammatiske kategoriar

Ordbøker viser grammatisk informasjon i ulik grad og på ulike måtar. Artiklar er merkte med ordklasse og eventuelt bøyingsmønster. Definisjonstekstar inneheld òg ofte grammatiske termar for å vise kva for samanhengar eit ord er brukt i. Endringar i grammatikk og grammatisk inndeling er relativt sjeldan, men ikkje uhøyrd. I 2005 kom Språkrådet og Utdanningsdirektoratet med ei anbefaling av nye grammatikktermar for skuleverket (Språkrådet 2005), som i stor grad baserte seg på *Norsk referansegrammatikk* (Faarlund et al. 1997). Inndelinga i referansegrammatikken skilde seg frå den gamle ved at ordklassene talord, artiklar og infinitivsmerke gjekk ut, og ordklassene subjunksjonar og determinativ kom til. Dei nye ordklassene tar i større grad omsyn til kva for utfylling orda tar. Tabell 1 viser gamle og nye ordklasser i referansegrammatikken og dei viktigaste endringane i den nye inndelinga.

TABELL 1. Gamal og ny inndeling av ordklasser og skilnadene.

Gamal inndeling	**Ny inndeling**	**Skilnad**
Substantiv	Substantiv	
Verb	Verb	
Adjektiv	Adjektiv	I tillegg: ordenstal
Pronomen	Pronomen	Endring: Fleire undergrupper av pronomen går til determinativ, og *som* går til subjunksjon og preposisjon
Artikkel		Går ut som ordklasse Artiklar går til determinativ
	Determinativ	Ny ordklasse Artiklar, nokre pronomen, grunntal
Talord		Går ut som ordklasse Grunntal går til determinativ, ordenstal går til adjektiv
Adverb	Adverb	
Preposisjon	Preposisjon	I tillegg: *som*, *enn* med substantivfraser og pronomenfraser
Interjeksjon	Interjeksjon	
Konjunksjon	Konjunksjon	Endring: Underordnande konjunksjonar (*at*, *om*, *fordi* osb.) går til subjunksjon
Infinitivsmerke		Går ut som ordklasse Infinitivsmerke (*å*) går til subjunksjon
	Subjunksjon	Ny ordklasse Omfattar infinitivsmerket og tidlegare underordnande konjunksjonar, *som* og *enn* med leddsetning som utfylling

Som offisielle og normgjevande held standardordbøkene seg til den gjeldande grammatiske terminologien. Standardordbøkene kom ut for første gang i 1986, og begge ordbøkene har blitt oppdaterte etter nye normer, særleg etter ny rettskriving for bokmål i 2005 og nynorsk i 2012, men det har ikkje vore løyvd pengar til ei omfattande innhaldsrevidering av ordbøkene. Etter at standardordbøkene kom på nett i 1994, har ein del nye ord og tydingar blitt lagde til, men Revisjonsprosjektet er den første gjennomgåande revisjonen. Vi kan no raskt endre informasjonen i artikkelhovudet med bruk av ordbanken, men det som er skrive i sjølve artiklane, må vi gå gjennom ein for ein. Det vil seie at skrivemåten og den grammatiske informasjonen i hovudet på artiklane er oppdatert, men der det er grammatisk informasjon i artiklane, nyttar desse inndelinga og terminologien frå det gamle systemet. Dette er tydeleg i dei meir komplekse funksjonsorda.

sia el. **sidan** adv (norr *síðan*; smh med *II sist*) **1** etter den tid, etterpå, seinare *det skal vi snakke om s- / s- fór dei heim* / så og så lang tid etter at noko hende *det er to år s-* / endeleg, omsider *seint og s-* **2** tidskonj: (i tida) etter at *du har vakse s- eg såg deg* **3** årsakskonj: fordi, på grunn av at *s- du kjem så seint, lyt du stå / han kan betale s- han har så god råd* **4** prep: frå – av; etter *ho har budd der s- nyttår*

FIGUR 3. Oppslaget for *sia/sidan* i 3. utgåve av Nynorskordboka, 2001, før den nye ordklasseinndelinga.

Eit eksempel er ordet *sia/sidan* (figur 3). I papirordboka frå 2001 står ordet som eit adverb, men med undertydingar som viser at det er brukt som tidskonjunksjon, årsakskonjunksjon og preposisjon. Konjunksjon er her brukt etter den gamle inndelinga, *sidan* blir i dag rekna som ein subjunksjon. Ettersom ein ikkje har gått inn i den einskilde artikkelen for å endre denne informasjonen, har han blitt ståande urørd fram til dagens revisjon. I den nye utgåva er artikkelen for *sia/sidan* splitta i tre homografar – adverb, subjunksjon og preposisjon – etter funksjon og ordklasse, òg for å gjere dei ulike funksjonane i ordet klårare. Dette følger òg trenden der digitale ordbøker, utan dei same plassavgrensingane som dei fysiske, i større grad gjev ulike funksjonar eigne oppslag. Denne

splittinga gjer det enklare for hovudbrukargruppa, som kjenner inndelinga frå lærebøker, å gjere seg nytte av den grammatiske informasjonen i artiklane.

Norsk Ordbok brukar den tradisjonelle ordklasseinndelinga. Ved kopling til ordbanken vil Metaordboka få tilgang til oppdaterte opplysingar om ordklasse, bøying og rettskriving.

5. Geografisk-administrative inndelingar

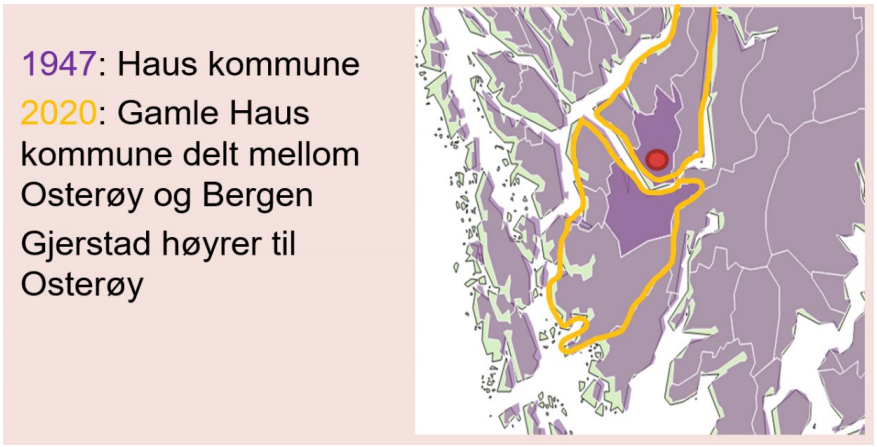
Noreg har hatt fleire større reformer der administrative einingar er slått saman. Resultatet er nye kommunegrenser og reduksjon av talet kommunar. I begge typane ordbøker gjer nye kommunar og kommunegrenser at ein må ta stilling til korleis ein skal handsame nemningar for innbyggarar i nye og gamle kommunar. For *Norsk Ordbok*, som inneheld dialektopplysingar og plasserer dialektord og -uttale geografisk, har slike endringar ført til at dei geografiske referansane i fleire tilfelle blir upresise.

Artiklane i *Norsk Ordbok* bygger på dokumenterte førekomstar av oppslagsord i skrift og tale. Dokumentasjonen av dialektord og -uttale kjem frå informantar i form av ordopplysingar frå setlar eller dialektordsamlingar. Opplysinga om kvar i landet ordet eller uttalen er henta frå, kallar ein *heimfesting*, og lågaste heimfestingsnivå er kommunen. NO baserer seg på kommuneinndelinga frå 1947 som var det tidspunktet landet hadde flest kommunar (747). Etter siste reform i 2020 er det 358 kommunar i landet. Mange av kommunane i 1947-inndelinga eksisterer dermed ikkje i dag, eller ser annleis ut enn tidlegare. Ordboka har digital kartvising, som gjer det mogeleg for brukarane å sjå kvar dei gamle kommunane låg.

I nettutgåva av NO blir opplysingar om dialektal variasjon i uttalen av ord viste i feltet Målføreformer. I figur 4, som gjev eit utsnitt av nettvisinga til artikkelen for substantivet *regn*, ser vi dialektopplysingane i det kvite feltet under oppslagsordet. Dette ordet har to registrerte uttalevariantar, *rign* og *ringn*. Uttalen er heimfesta til Nissedal kommune på Austlandet, Agder fylke på Sørlandet, og Sogn (So), Jostedal og Sunnfjord (Sfj), som alle er del av det som tidlegare var Sogn og Fjordane fylke.

FIGUR 4. Dialektopplysingar for substantivet *regn* i Norsk Ordbok.

Eit eksempel på unøyaktigheit som følger av nye kommuneinndelingar, er heimfestingar frå den tidlegare kommunen Haus, som låg på begge sider av Osterfjorden, i det som då var Hordaland fylke. Ein av informantane til *Norsk Ordbok* kjem frå Gjerstad sør på øya Osterøy. I 1964 blei Haus og fleire andre kommunar slått saman til Osterøy. Figur 5 viser korleis gamle Haus kommune (mørk lilla) etter 1964 er delt mellom det som no er kommunane Osterøy og Bergen (gule omriss) på kvar si side av fjorden – Osterøy på nordsida og Bergen på sørsida. Gjerstad (raud sirkel) høyrer til dagens Osterøy.



FIGUR 5. Gamle Haus kommune er i dag delt mellom Osterøy og Bergen.

Når ei uttaleopplysing i NO er heimfesta til Haus kommune, kan dette altså referere anten til ein del av dagens Bergen kommune eller ein del av dagens Osterøy kommune. Om ein går inn i det digitale setelarkivet og søker på Haus, vil ein sjå at dei fleste opplysingane derfrå kjem frå informanten frå Gjerstad, og dermed frå Osterøy. Men i kartvisinga, som viser 1947-inndelinga med kommunen som lågaste geografiske nivå, får ein ikkje ei presis lokalisering. Der får ein berre sjå kommunen Haus – som altså låg på begge sider av fjorden i 1947 (jf. figur 5). Med tilgang til kartvising og kjeldemateriale kan brukarane sjølve undersøke opplysingane ved å gå til kjeldematerialet.

Å bevare den opphavelege geografiske inndelinga i NO har i seg sjølv dokumentasjonsverdi. Relativt få av dagens brukarar kjenner 1947-inndelinga, og vi må rekne med at dei både vil ønske og stille krav om å kunne søke etter dagens geografiske inndelingar. I kartvisinga kan ein sjå føre seg fleire løysingar når ei dialektopplysing er heimfesta til Haus:

1. Vise ny inndeling, dvs. både Osterøy og Bergen, sidan Haus var delt mellom det som i dag er to kommunar. Til saman dekker dei eit mykje større område enn Haus gjorde, og ei slik heimfesting blir dermed endå meir upresis enn ho er i dag.
2. Vise dagens Osterøy kommune. Då bør redaksjonen gå gjennom alle gongane Haus er brukt i *Norsk Ordbok*, og kontrollere om belegga faktisk er frå Osterøy.
3. Vise geolokalisering (punktmerking), som er uavhengig av administrative inndelingar, slik ein kan sjå det i f.eks. Google Maps.
4. La brukarane velje mellom ny og gammal inndeling når dei søker, og sørge for god dokumentasjon og bruksrettleiing.

Felles for *Norsk Ordbok* og standardordbøkene er spørsmålet om korleis vi skal presentere kommunale innbyggarnemningar. Kva gjer vi med utdaterte nemningar for innbyggarar i nedlagde kommunar, og kva gjer vi når nye namn enno ikkje er bestemt eller har festa seg i bruk?

Standardordbøkene viser berre nemningar for innbyggarar som er i faktisk, allmenn bruk. For å unngå formuleringar som «i tidlegare x kommune» og dermed knytte ordet til administrative einingar, nyttar standardordbøkene etablerte namn på geografiske regionar, som stadnamna Jæren eller Haugalandet framfor det administrative namnet Rogaland,

sidan dei geografiske namna er meir presise og mindre omskiftelege enn dei administrative. Innbyggarnemningane i standardordbøkene blei gjennomgått systematisk hausten 2019, før den siste kommunereforma i 2020, og det er fleire nemningar frå kommunar som ikkje lenger finst i ordbøkene.no, som *lindåsing*, etter den tidlegare kommunen Lindås. Denne kommunen er no ein del av den nye kommunen Alver. Den nye nemninga *alverbu* har enno ikkje festa seg heilt i språket, og er per dags dato ikkje tatt inn i standardordbøkene, men kan takast inn om bruken held fram.

NO-AH har kartlagt nye og nedlagde kommunar i alfabetstrekket *a–h* og innbyggarnemningar knytt til dei aktuelle kommunane. Sidan NO brukar inndelinga frå 1947, tar ein med innbyggarnemningar frå denne perioden. Problemstillinga med nedlagde kommunar gjeld også kommunar som blei lagde ned i 1967, på same måte som kommunar som forsvann i 2020. Det viser seg også at seinare redigeringspraksis ikkje har halde seg strengt til 1947-inndelinga. På bakgrunn av undersøkinga blir retningslinjene for innbyggarnemningar no reviderte.

6. Oppsummering og diskusjon

I en revisjonsprosess må ein forhalde seg til endringar. Vi har prøvd å vise ulike utfordringar knytt til endringar i rettskriving, grammatikk og geografi i standardordbøkene og *Norsk Ordbok*, og korleis vi har prioritert og prøvd å løyse desse utfordringane på bakgrunn av typen og brukarane av den aktuelle ordboka.

At norsk rettskrivingspolitikk er prega av stor valfridom og frekvente justeringar, fører med seg krav om fortløpande oppdatering av ordbøker som skal vise norma. I begge prosjekt blir endringar i rettskrivinga handterte ved at ordbøkene er kopla til digitale rettskrivingsressursar. Drift og vedlikehald av rettskrivingsdatabasane er dermed avgjerande for at ordbøkene skal kunne vise fram gjeldande rettskriving og grammatikk. Det viktigaste i standardordbøkene er å vise fram den gjeldande norma, og koplinga til ordbanken gjer at artikkelhovudet alltid er oppdatert. I NO-AH blir visning av gjeldande rettskriving og nye grammatiske kategoriar mogeleg ved oppdatering av Metaordboka og kopling av Metaordboka til ordbanken.

At *Norsk Ordbok* brukar ei eldre geografisk inndeling kan komme i konflikt med brukarane sine forventningar og krav. Vi må derfor vurdere

om vi skal vise fram, og gjere det mogeleg å søke etter, ny og gammal inndeling i heimfesting og kartvisning. I begge prosjekta må redaksjonane vere merksame på at kommunegrenser endrar seg, men nemningar for innbyggarar i nye kommunar må feste seg i bruk før ein kan ta dei inn i ordbøkene.

Mange tenker på nettordbøker som ein del av nettet, og at desse er underlagde dei same reglane for søk som søkemotorar. Både standardordbøkene og *Norsk Ordbok* har behov som kan løysast med hjelp av digitale løysingar, men brukarane har ulik digital kompetanse, ulike forventingar og stiller ulike krav.

Brukarundersøkingane på nettsidene til ordbøkene.no viste at viktigaste grunnen til at folk søker i ordbøkene, er rettskriving (22,4 %) og bøying (32,4 %), og det er å vente at brukarane faktisk er usikre på skrivemåten. Dermed er det viktig at søkefunksjonen er god nok til at ein finn det aktuelle ordet, både i bøygd form og når det er feil stava, f.eks. ved hjelp av ein god «meinte du?»-funksjon. I eit språk som har såpass frekvente rettskrivingsendringar som norsk, kunne det i ei rettskrivingsordbok òg vere nyttig å knyte unormerte og tidlegare normerte former til dei normerte, slik at dei normerte formene blir vist når ein søker på dei ikkje-normerte. Dette er mogleg ved koplinga til ordbanken sidan informasjonen finst der.

Ein viktig del av den tekniske oppgraderinga av *Norsk Ordbok* blir å avklare kva for informasjon ein ønsker å gjere tilgjengeleg for brukarane i vising og søk, og korleis det skal gjerast. Grunnlagsmaterialet til NO er mykje meir innhaldsrikt enn det som er vist i dagens ordbok. Med nye tekniske løysingar vil det bli mogeleg å vise fram variasjon i skrift og tale i større grad. Ein kan presentere geografisk informasjon på ulike måtar, og frå ulike epokar. Dei geografiske opplysingane i ordbokbasen gjer det i teorien mogeleg å søke etter ord og uttale ikkje berre frå tidlegare kommunar, men også frå ein særskild stad. Her kan også ein «meinte du»-funksjon brukast til å hente fram dialektvariantar, noko som ikkje er mogeleg i dag.

Digitalt format gjer det mogeleg å integrere og tilpasse informasjon, og i større grad la brukarane velje sjølve. Med det meiner vi at vi kan gje brukarane val, slik at ein kan søke både i nye og gamle geografiske inndelingar eller i normerte og unormerte former, og likevel få treff som blir oppfatta som relevante og riktige. Det digitale formatet gjev oss høve til å ikkje berre lage ei ordbok for éi tenkt målgruppe, men å femne alle – ikkje berre dagens, men også morgondagens brukarar.

Litteratur

Faarlund, Jan Terje, Svein Lie & Kjell Ivar Vannebo 1997. *Norsk referansegrammatikk*. Oslo: Universitetsforlaget.

Grønvik, Oddrun & Christian-Emil Ore 2014. Samvirket mellom ordbank og ordbok. I: Fjeld, Ruth Vatvedt & Hovdenak, Marit (red.). *Nordiske studier i leksikografi 12*. Oslo: Nordisk förening för lexikografi, 139–158.

Grønvik, Oddrun & Christian-Emil Ore 2018. Bokmål og nynorsk samindeksert – Metaordboka som verktøy for jamføring og utforsking av ordtilfang. I: Svavarsdóttir, Ásta, Halldóra Jónsdóttir, Helga Hilmisdóttir & Þórdís Úlfarsdóttir (red.). *Nordiske studier i leksikografi 14*. Reykjavik: Nordisk förening for lexikografi, 87–95.

Hovdenakk, Marit, Laurits Killingbergtrø, Arne Lauvhjell, Sigurd Nordlie, Magne Rommetveit & Dagfinn Worren (red.) 2001. *Nynorskordboka*. 3. utgåve. Oslo: Samlaget.

NO = Norsk Ordbok. 1930–. Oslo: Samlaget. Universitetet i Bergen. I: <no2014.uib.no> og <norsk-ordbok.no>. Henta september 2022.

Ordbanken = *Norsk ordbank – nynorsk* <https://inger.uib.no/perl/search/search.cgi?appid=73&tabid=1116> *Norsk ordbank – bokmål* <https://inger.uib.no/perl/search/search.cgi?appid=72&tabid=1106>. Henta februar 2023.

ordbøkene.no = *Bokmålsordboka*. Språkrådet og Universitetet i Bergen & *Nynorskordboka*. Språkrådet og Universitetet i Bergen. I: <ordbøkene.no>. Henta september 2022.

Ore, Christian-Emil 2000. Metaordboken – et rammeverk for Norsk Ordbok. I: Gellerstam, Martin, Kristinn Jóhannesson, Bo Ralph & Lena Rogström (red.). *Nordiska studier i lexikografi 5*. Göteborg: Nordisk förening för lexikografi 250–270.

Rauset, Margunn 2022. Brukarmedverknad i utvikling av nettsida ordbøkene.no. *LexicoNordica* 29, 97–118.

Språkrådet og Utdanningsdirektoratet 2005. *Grammatiske termer til bruk i skoleverket*. <https://www.sprakradet.no/globalassets/sprakhjelp/gramterm.pdf>.

Vikør, Lars 2018. *Inn i Norsk Ordbok. Brukarrettleiing og dokumentasjon*. Oslo: Samlaget.

”Varför står det olika i SAOL och i SO?” Om (bearbetning av) skillnader mellan Svenska Akademiens samtidsordböcker

Emma Sköldberg

The Swedish Academy's contemporary dictionaries, the glossary SAOL and the definition dictionary SO, have many features in common but they also show a lot of differences, especially in terms of content. Many of these differences can be explained with reference to the perspectives, traditions and publication year of the dictionaries. However, some differences are difficult to justify. For this reason, the editorial team of SAOL and SO is currently working on 1) identifying and 2) managing differences with respect to the information given in the two lexical resources. In this article, I discuss different types of differences, both motivated and unmotivated, between the dictionaries. The issue of priorities in the editorial work concerning unmotivated differences between SAOL and SO is also addressed.

NYCKELORD: lexikografi, ordlista, definitionsordbok, harmonisering, informationskategori

1. Inledning

De två samtidsordböckerna, *Svenska Akademiens ordlista* (SAOL, 14 uppl., 2015) och *Svensk ordbok utgiven av Svenska Akademien* (SO, 2 uppl., 2021), är sedan några år tillbaka tillgängliga via ordboksportalen www.svenska.se. I portalens användargränssnitt söker ordboksanvändarna i de båda verken, samt historiska *Svenska Akademiens ordbok* (SAOB 1898–), på samma gång. De tre ordböckerna kompletterar varandra på flera sätt och av detta skäl har visningssättet många fördelar. Vidare kan presentationssättet leda till att användare, som söker efter information i ett av verken, uppmärksammas på att det finns andra ordböcker som kan vara lämpligare med tanke på de upplysningar som hen är i behov av (se t.ex. Lönnroth 2018; Bäckerud et al. 2020). Men visningssättet leder också till användarfrågor gällande uppgifter som inte är samstämmiga i samtidsordböckerna (se vidare avsnitt 2 nedan). Med tanke på att Svenska

Akademien står bakom samtliga verk i portalen och att både SAOL och SO gäller modern svenska är det begripligt att dessa frågor infinner sig.

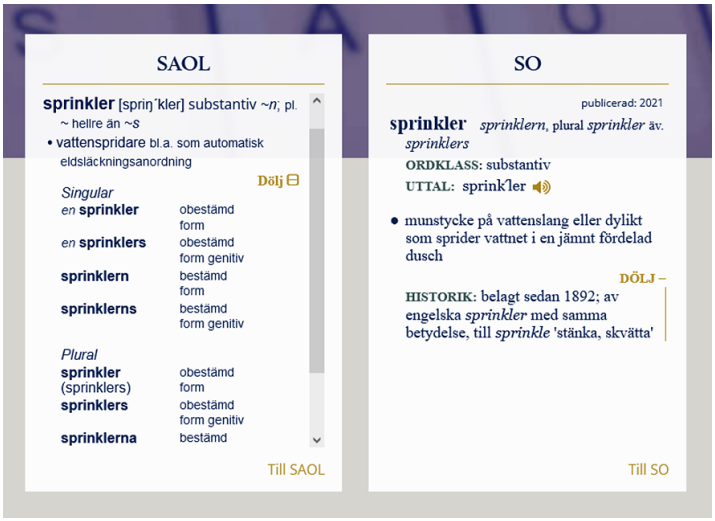
I det följande redogör jag för olika typer av skillnader mellan samtidsordböckerna SAOL och SO. Vidare resonerar jag kring ett pågående arbete med att, inför kommande upplagor, harmonisera motstridiga uppgifter i de båda verken. Exempel hämtas från olika informationskategorier. Men först några ord om de två ordböckerna och bakgrunden till det pågående arbetet.

2. Akademiens två samtidsordböcker

SAOL och SO har många gemensamma drag. Båda ordböckerna är enspråkiga och gäller svenskt samtidsspråk. De har också många gemensamma ord i sina respektive lemmalistor.

I det här bidraget fokuserar jag emellertid på skillnaderna mellan verken. En övergripande skillnad är att SAOL och SO utgör olika typer av lexikografiska verk och att de har olika perspektiv (se vidare Svensén 2004:26–47). Medan SAOL är en mer normativ ordlista vars främsta syften är att stödja stavning och böjning av svenska ord är SO en mer deskriptiv definitionsordbok med tyngdpunkt på ordens betydelse och användning (jfr beskrivningarna av *ordbok*, *definisjonsordbok* och *ordliste* i Bergenholtz et al. 1997). SAOL stöder därmed i första hand produktion. SO stöder främst reception men också produktion, t.ex. genom uppgifter om uppslagsordens uttal, synonyma uttryckssätt, fraseologi och konstruktionsmönster (se vidare Malmgren 2009). En viktig skillnad är också att ordböckerna, som redan antytts, har olika publiceringsår: den fjortonde upplagan av SAOL blev tillgänglig sex år tidigare än den andra upplagan av SO.

I samtidsordböckernas ordboksartiklar finns fler skillnader, inte minst i fråga om de uppgifter som ges. Som exempel kan artikeln **sprinkler** i de båda verken nämnas (se figur 1). SAOL-artikeln återfinns till vänster och SO-artikeln till höger.



FIGUR 1. Ordboksartiklarna **sprinkler** i SAOL och SO i svenska.se (något beskurna).

Som framgår av figur 1 innehåller SO, till skillnad från SAOL, bl.a. ljudfiler som kompletterar de skrivna uttalsangivelserna. Vidare tillhandahåller SO historiska uppgifter om orden. I SAOL-artiklarna finns det däremot fylligare uppgifter om ordens böjningsmönster. Gällande böjningsuppgifterna kan man också notera att det ibland, som vid *sprinkler*, är olika uppgifter som presenteras. I SAOL står det följande om ordet pluralform: ”*sprinkler* hellre än *sprinklers*”. I SO står det i stället ”*sprinkler* äv. *sprinklers*”. Orsaken till denna skillnad torde vara att ordlistan, som redan nämnts, är normativ och att man inom svensk språkvård av tradition har motarbetat engelskt s-plural (se vidare Malmgren 2014:83; se även avsnitt 3.7 nedan).

Det har alltid funnits skillnader mellan de uppgifter som tillhandahålls av SAOL och SO, men genom publiceringen i svenska.se är dessa skillnader mer uppenbara för dagens ordboksanvändare. Iakttagelser av skillnader leder inte sällan till mejl till ordboksredaktionen med frågor av typen ”Vad har egentligen *både* för ordklass? I SAOL är ordet en konjunktion eller ett substantiv och i SO är det ett adverb eller ett substantiv”. Orsaken till just denna uppgiftsskillnad är att SO-redaktionen har valt att anamma den ordklassanalys av ordet som finns i *Svenska Akademiens grammatik* (SAG, 1999). Allt talar för att SAOL-uppgifterna om ordet *både* kommer att ha ändrats i nästa upplaga av ordlistan och att ordklassuppgifterna i ordböckerna och i grammatikan därmed blir enstämmiga. Just denna

skillnad mellan SAOL och SO beror alltså på att ordboksartiklarna har författats vid olika tidpunkter och att innehållet i verken inte är helt i takt.

Ett annat exempel rör substantivet *tillstånd* och dess båda betydelser ’situation, läge’ respektive ’tillåtelse, medgivande’. Mejlfrågan lyder ”Varför kommer substantivet *tillstånd*:s båda betydelser i olika ordning i de båda verken? Har det blivit fel?” I detta fall är det inte självklart vilken av de två fullt levande betydelserna hos *tillstånd* som ska placeras överst.[1] Orsaken till just denna skillnad mellan verken torde vara att ordböckerna har olika bakgrund och att de har utarbetats inom delvis olika miljöer (se vidare om SAOL:s historia i bl.a. Gellerstam 2009a:53–83 och om SO:s bakgrund i Malmgren & Sköldberg 2013; se även Josephson 2022:254–257).

Att ett och samma ord har olika ordklassuppgifter eller att betydelser kommer i olika ordning i SAOL respektive i SO utgör inget problem för de användare som konsulterar en ordbok i taget. Däremot kan olikheter som dessa, som synes, förvirra ordboksanvändare som tittar i svenska.se och leda till tolkningsproblem.

3. Skillnader mellan SAOL och SO – utifrån olika informationskategorier

Som redan framkommit kan till synes omotiverade uppgiftsskillnader mellan SAOL och SO inverka menligt på användarnas förståelse av de uppgifter som ges. Av detta skäl pågår det nu ett större arbete inom ordboksprojektet med att bedöma och eventuellt bearbeta skillnader kopplade till de upplysningar som ges i Akademiens samtidsordböcker. En förutsättning för detta arbete är givetvis att redaktionen har en god bild av vad skillnaderna faktiskt består i.

I det följande kommer jag, utifrån olika informationskategorier, att lite mer ingående diskutera typer av skillnader mellan verken. Med tanke på att det totala ordboksmaterialet är mycket omfattande kommer jag

1 När SAOL-ordet försågs med betydelser (i upplaga 9, 1950) placerades betydelsen ’tillåtelse’ överst. Under flera efterföljande SAOL-upplagor förklarades uppslagsordet inte, men sedan upplaga 14 (2015) är de semantiska uppgifterna tillbaka och betydelsen ’situation’ står nu först. Ändringen kan ha att göra med hur vanliga de båda betydelserna är i olika sammansättningar. I SO har betydelsen ’tillåtelse’ förmodligen placerats överst i artikeln eftersom den är något äldre.

endast att göra vissa nedslag. Vidare fokuserar jag på informationskategorier som är gemensamma för SAOL och SO och bortser från sådana som är unika för något av verken, t.ex. etymologiuppgifter. Sådana finns endast i SO.

3.1. Lemmaurval

En viktig skillnad mellan SAOL och SO är att SAOL innehåller ungefär dubbelt så många uppslagsord som SO. SAOL och SO listar dessutom delvis olika typer av uppslagsord. Ordlistan upptar t.ex. fler (mer) genomskinliga sammansättningar och den behandlar också, till skillnad från SO, egennamn (såsom **Tanzania** och **Stalin**) (se vidare Malmgren 2014:95–96). Därutöver innehåller ordlistan, sedan upplaga 11 (1986), fler finlandssvenska ord (t.ex. **geronom** och **memma**). Enligt Gellerstam (2009a:76–77) beror det på att dessa ord, i kraft av finlandssvenskans status som officiellt språk i Finland, intar en särställning i förhållande till andra regionala ord. SO har å andra sidan verbförbindelser i form av partikelverb, reflexiva verb och reflexiva partikelverb bland sina uppslagsord (se t.ex. **duka under**, **fåna sig** och **ta igen sig**). Sådana verbförbindelser utgör, i den mån de behandlas, uppslagsord ”av andra ordningen” i den fjortonde upplagan av SAOL (se SAOL 2015:XIX; se även Malmgren 2014:93–94).

3.2. Lemmalösning

I normalfallet är det den oböjda formen av ett ord som utgör uppslagsformen i en ordbok. Denna regel har dock sina undantag. Om exempelvis ett substantiv bara används i plural bör, enligt Svensén (2004:128), pluralformen utgöra lemmaform (t.ex. *inälvor*). Används ett substantiv nästan uteslutande i bestämd form bör i stället bestämd form utgöra uppslagsform (t.ex. *tyngdlagen*).

En granskning av de båda lemmalistorna visar att jämförbara uppslagsord huvudsakligen har samma form i SAOL och i SO men att det också finns skillnader. Exempelvis utgör ordet **seglarstövel** uppslagsord i SAOL. I SO är motsvarande lemma **seglarstövlar**. I SAOL återfinns uppslagsordet **geniknölar** och i SO lemmat **geniknölarna**. Det är svårt att säga varför det har blivit så här, men en sak står klar: olikheter som dessa kan uppfattas som inkonsekvenser av användarna.

3.3. Homografi och lemmaordning

Ytterligare en skillnad gällande SAOL:s och SO:s lemmalistor består i den ordning som vissa homografa ord presenteras i. Närmare bestämt finns det fall där båda verken listar samma homografa ord men att dessa behandlas i inbördes olika ordning. Exempelvis listas **cash** först som substantiv och sedan som adverb i SAOL. I SO är ordningen den motsatta. Det finns inget i SAOL:s eller SO:s redaktörsanvisningar som reglerar i vilken ordningen homografa ord tillhörande olika ordklasser ska placeras. Skillnaden i ordning kan däremot bero på när de olika uppslagsorden införlivats i respektive verk.[2] Ett annat exempel är **köra** (med betydelserna ’sjunga i kör’ och ’styra fordon’) som utgör två verb i respektive verk. Också dessa har olika ordningsföljd i SAOL och SO. Även denna typ av olikhet är svår att förklara för användarna och den komplicerar en jämförelse mellan de uppgifter som ges i respektive resurs.

3.4. Stavning och stavningsvariation

Enligt Svensén (2004:134) torde uppgifter om stavning vara en av de viktigaste informationstyperna i en allmänordbok. Stavning är också en av de informationskategorier som oftast är föremål för språkvårds- och normeringsverksamhet.

En granskning av SAOL:s och SO:s lemmalistor visar att huvuddelen av deras uppslagsord stavas på ett och samma sätt. Men det finns en hel del undantag också från denna regel. Skillnaderna rör bl.a. 1) vilken uppslagsform det är som har valts ut, 2) om det presenteras stavningsvarianter eller ej, samt 3) eventuella stavningsvarianters inbördes ordning och styrkeförhållanden. Ett första exempel är SAOL-lemmat **k-sprit** vars motsvarighet i SO är **K-sprit**. En uppmärksam användare kan givetvis undra varför det är så. Ett annat fall är SAOL:s uppslagsord **kasino**. I SO återfinns också denna form men även en stavningsvariant. Det står ”**casino** eller **kasino**”. Skillnaderna gällande uppslagsformer och varianternas inbördes ordning kan ofta förklaras med att SAOL är norme-

2 I SAOL lades adverbet **cash** till senare än det likalydande substantivet. I SO har de två orden haft denna ordningsföljd sedan de båda inkluderades i föregångaren till SO, *Svensk ordbok* (SOB, 1986).

rande och SO beskrivande. Stavningsvarianten *casino* är mer frekvent än *kasino* i t.ex. de korpusar som SO-redaktionen konsulterat under arbetets gång.

I SAOL används, till skillnad från i SO, också notationssättet ”hellre än” mellan två uppslagsformer. Ordkombinationen används dels när den andra stavningsvarianten är påtagligt svagare i bruket än förstaformen (som vid ”**sjal** hellre än **schal**”), dels när en viss stavning, av språkvårdsskäl, förordas framför en annan (t.ex. ”**sprej** hellre än **spray**”; se även figur 1 ovan) (SAOL 2015:XVIII). I deskriptiva SO används antingen ”eller” eller ”även” mellan alternativformerna. När ”eller” används betraktas formerna mer eller mindre som likvärdiga. När ”även” används är alternativformen exempelvis lite ovanligare. I just dessa fall står det ”**sjal** även **schal**” och ”**spray** eller **sprej**” i SO.

3.5. Ordklass

Ordklassangivelsen utgör en central del av den grammatiska information som ges om uppslagsorden (se vidare Holmer 2016:39–40 och där anförda referenser). Det är emellertid inte självklart vilken ordklassuppsättning det är som ska användas i ett verk. Uppsättningen kan också revideras mellan olika upplagor av samma ordbok (se vidare Svensén 2004:179–185; 450; se även Holmer 2016:62–65 om ordklassangivelser i olika upplagor av SAOL). Som exempel infördes den nya ordklassen *subjunktion* i SO (2009) och i den fjortonde upplagan av SAOL (2015). På så sätt anslöt sig SO och SAOL (åtminstone i denna fråga) till terminologin i SAG (1999).

SAOL:s och SO:s ordklassuppsättningar är inte identiska. Som framgick av exemplet **både** kan själva ordklassuppgiften gällande ett enskilt uppslagsord också skilja sig åt mellan ordböckerna (se avsnitt 2). I SAOL finns det även enstaka uppslagsord, som **korsvis**, vilka försetts med dubbel ordklassbeteckning, här adverb och adjektiv. I den mån dessa homografa ord listas i SO bildar de två separata ordboksartiklar. Vid en sökning på den typen av homografer i svenska.se tycks de innehållsmässiga skillnaderna mellan verken vara större än de egentligen är.

3.6. Uttal

En jämförelse mellan SAOL:s och SO:s uttalsuppgifter visar att det, utöver att endast SO tillhandahåller inläst uttal, för det första finns skillnader som rör om det över huvud taget finns en uttalsuppgift eller ej. I båda ordböckerna förväntas användarna känna till vissa allmänna uttalsregler när det gäller svenska språket. Men utöver detta förväntas SAOL-användaren kunna dra slutsatser om uttal genom att titta på hur andra ord i samma verk uttalas. Exempel är uppslagsorden **bowling** och **bowlinghall** som har uttalsangivelse i SO men som saknar uttalsangivelse i SAOL. I ordlistan måste användaren titta på hur ordet **bowla** uttalas och dra slutsatser utifrån det.

För det andra finns det skillnader i hur ett och samma uttal anges, dvs. vilka notationsprinciper som nyttjas (jfr t.ex. **reggae** [reg´ej] i SAOL med **reggae** [reg´ei] i SO) (se vidare t.ex. Svensén 2004:144–147 om typer av uttalsnotation).

För det tredje finns det skillnader rörande vilket eller vilka uttal uppslagsorden anges ha. Exempel är uppslagsordet **router,** som åtföljs av uttalsuppgiften [ro´ter el. ra^{o}´ter] i SAOL men enbart [ra^{o}´ter] i SO.

För det fjärde kan uttalsvarianterna komma i olika ordning. Se t.ex. **kaviar** med notationen ”[-a´r el. kav´i]” i SAOL och notationen ”kav´iar el. kavia´r” i SO.

3.7. Böjning

En av SAOL:s huvuduppgifter är att beskriva och normera svensk ordböjning. Detta gäller inte minst för den fjortonde upplagan i vilken alla substantiv, verb och adjektiv, och då även sammansättningar, har försetts med böjningsangivelser (SAOL 14, 2015:IX). SAOL ger även långt fler böjningsformer för varje uppslagsord än vad SO gör. I exempelvis substantivartikeln **äpple** ges sammanlagt åtta former i ordlistan och tre former i ordboken. Vid ett verb som **vabba** ges totalt tio former i SAOL och tre i SO.

I normativa SAOL förses uppslagsorden så långt det är möjligt med svenska böjningsformer (Gellerstam 2009b:30) och det resulterar ibland i att ordlistan presenterar andra böjningsformer än vad deskriptiva SO gör. Ett exempel är redan nämnda *sprinkler* (se figur 1 ovan). Ett annat exem-

pel är substantivet *voucher*. I SAOL böjs ordet *voucher*, *vouchern*, plur. *vouchrar*. I SO böjs det *voucher*, *vouchern*, plural *vouchrar* el. *voucher* äv. *vouchers*. Det beror på att det finns tre relativt frekventa pluralformer av detta ord i de korpusar som SO-lexikograferna konsulterat.

3.8. Betydelseangivelse

I en definitionsordbok som SO är ordförklaringarna centrala. I SAOL är de semantiska uppgifterna relativt begränsade även om det har blivit fler betydelseangivelser i upplaga 14 än i tidigare upplagor (se t.ex. Malmgren 2014:91–93). Trots denna förändring i SAOL är skillnaderna mellan verken på denna punkt mycket stor. Som exempel kan verbartikeln **ta** nämnas. I SAOL saknas en betydelseangivelse. Användarna förväntas känna till vad detta ord betyder. I SO har ordet inte mindre än 17 huvudbetydelser och ett mycket stort antal underbetydelser.

Ett annat exempel är substantivet **mandolin** som anges vara ’ett stränginstrument’ i SAOL. Motsvarande SO-artikel redovisar samma betydelse men också en nyare sådan, dvs. ’ett köksredskap som används för att göra tunna skivor och strimlor’. Här har SO från 2021 inkluderat en nyare betydelse som ännu inte har införlivats i SAOL.

Ett tredje exempel är artikeln **kabellängd**. I SAOL förklaras inte ordet men i artikeln finns det två hänvisningar, en till uppslagsordet **kabel** och till **längd 1**. I SO har ordet följande definition: ’en tiondels nautisk mil dvs. ca 185 meter’. Ordet *kabellängd* är alltså inte en genomskinlig sammansättning och i ett fall som detta kan hänvisningarna i ordlistan betraktas som missvisande (se vidare Blensenius et al. 2021).

3.9. Brukskommentar

Ytterligare en skillnad mellan verken rör deras brukskommentarer, dvs. kommentarer om uppslagsordens stil, värdeladdning, bruklighet m.m. För det första ser inte uppsättningarna med kommentarer likadana ut. I SAOL används t.ex. kommentaren ”prov.” (provinsiellt) vid ett ord som **bamba**. I den andra upplagan av SO har ”prov.” ersatts av mer begripliga ”dialektalt”. För det andra kan ett och samma uppslagsord behandlas olika när det gäller just förekomst av kommentar. Exempelvis har det mer talspråkliga uppslagsordet **kaffegök** kommentaren <vard.> i SAOL

men ingen kommentar i SO. Ett annat exempel är **indian** med kommentaren <kan uppfattas som nedsättande> i SO. Samma ord är omarkerat i SAOL-upplagan från 2015. Exemplet illustrerar att språkbrukarnas syn på ett visst ord kan förändras relativt snabbt. Det är därför fördelaktigt med löpande uppdateringar av lexikografiska verk.

4. Motiverade och omotiverade skillnader mellan samtidsordböckerna

De skillnader som finns mellan uppgifterna i Svenska Akademiens samtidsordböcker kan betraktas som mer eller mindre motiverade. En skillnad som går att motivera kan t.ex. vara att ett ord som *sprinkler* har olika böjningsangivelser i normativa SAOL och i deskriptiva SO. En annan kan vara att stavningsvarianterna *casino* och *kasino* kommer i olika ordning i de båda verken. Ännu ett exempel på en motiverad skillnad är att SAOL ger fler böjningsformer än SO. Detta beror på att SAOL i högre grad än SO avser att stödja produktion.

Andra skillnader kan däremot svårligen försvaras med hänvisning till att SAOL och SO utgör olika typer av verk, att de har olika perspektiv och att de avser att stödja olika funktioner. I denna artikel har jag gett flera exempel på sådana fall. Ett av dessa rör presentationen av substantivet *seglarstövel* vilket har singular form i SAOL och plural form i SO. Här, och i fall som dessa, borde ordboksredaktionen, för att förenkla för ordboksanvändarna, bestämma sig för en gemensam uppslagsform i verken. Ett annat exempel gäller ordningen på homograferna **kippa** (verb) och **kippa** (substantiv). Utan tvekan vore det tydligare i svenska.se om dessa ord kom i samma ordning i samtidsordböckerna.

Samtidigt finns det gränsfall, bl.a. när det gäller lemmaurvalet, som inte på samma sätt styrs av t.ex. ordbokstyp. Man kan exempelvis ifrågasätta att finlandssvenska ord hanteras på olika sätt i verken. Ett annat exempel är hanteringen av verbförbindelser i lemmalistorna, något som kan kopplas till de båda ordböckernas skiftande bakgrund. Det finns med andra ord anledning att diskutera och kanske ompröva tidigare beslut gällande såväl lemmaselektion som lemmaansättning i respektive verk.

Ännu ett exempel rör ordböckernas uttalsuppgifter. Att SAOL, till skillnad från SO, saknar uttal på vissa uppslagsord kan bero på att en ordlista typiskt tillhandahåller mindre information om varje uppslagsord och att

det historiskt har funnits begränsat med utrymme i de tryckta upplagorna. I en digital SAOL i svenska.se är detta knapphändiga presentationssätt betydligt svårare att försvara.

Man kan också diskutera de skillnader som rör betydelseangivelserna. Som redan nämnts förklaras inte ordet *ta* i SAOL. Andra ord som inte förklaras i ordlistan, men som definieras i SO, är t.ex. *stol*, *vacker* och *skratta*. Något förenklat kan man säga att SAOL, som just är en ordlista, inte förklarar de uppslagsord som användarna förväntas känna till, men frågan är hur man avgör vilka dessa ord är. Genom publiceringen på svenska.se utgör SAOL-användarna en mer heterogen grupp än tidigare. Exempelvis torde antalet svenskinlärare ha ökat bland användarna. Å andra sidan kan de SAOL-användare som har behov av förklaringar, hitta sådana i grannordböckerna i portalen.

5. Slutord

SAOL och SO, två verk med olika historia men numera med samma avsändare, visas upp parallellt i ordboksportalen svenska.se. Det finns, och har alltid funnits, stora skillnader vad gäller enskilda uppgifter i ordböckerna, men dessa olikheter var inte så tydliga när ordböckerna publicerades på olika håll.

Att mer exakt klarlägga vad skillnaderna mellan verken består i är ett omfattande arbete. Det kommer också alltid att finnas olikheter mellan dem så länge ordböckernas respektive upplagor inte publiceras samtidigt.

Ur användarsynpunkt är det dock angeläget att minska ner på antalet omotiverade skillnader mellan samtidsordböckerna. SAOL- och SO-redaktionens medlemmar måste hitta en lagom ambitionsnivå när det gäller det pågående harmoniseringsarbetet. I nuläget fokuserar redaktionen i första hand på mer formella olikheter, men också på sådana skillnader som rör presentationssätt (såsom vid homografa uppslagsord). Samtidigt måste medarbetarna fundera över vilka ändringar som är relevanta på sikt. Ordboksredaktörerna kan också behöva lyfta blicken ytterligare och fundera över likheter och skillnader (såväl motiverade som omotiverade) mellan samtidsordböckerna och SAOB och hur dessa tre ordböcker ska förhålla sig till varandra i framtiden.

Referenser

Bergenholtz, Henning, Ilse Cantell, Ruth Vatvedt Fjeld, Dag Gundersen, Jón Hilmar Jónsson & Bo Svensén 1997. *Nordisk leksikografisk ordbok*. Oslo: Universitetsforlaget.

Blensenius, Kristian, Louise Holmer & Emma Sköldberg 2021. SAOL 14 som rättesnöre – diskussion kring den senaste upplagan. *LexicoNordica* 28, 39–58.

Bäckerud, Erik, Pär Nilsson & Emma Sköldberg 2020. Så används Svenska Akademiens ordböcker på nätet. Implicit och explicit feedback från användarna. I: Caroline Sandström et al. (red.), *Nordiske studier i lexikografi 15*. Helsingfors, 91–100.

Gellerstam, Martin 2009a. SAOL i många upplagor. I: Martin Gellerstam (red.), *SAOL och tidens flykt. Några nedslag i ordlistans historia*. Stockholm: Norstedts, 53–83.

Gellerstam, Martin 2009b. Vad är Svenska Akademiens ordlista? I: Martin Gellerstam (red.), *SAOL och tidens flykt. Några nedslag i ordlistans historia*. Stockholm: Norstedts, 11–30.

Holmer, Louise 2016. *Grammatik i SAOL. En undersökning av grammatisk information i Svenska Akademiens ordlista över 130 år.* (Meddelanden från Institutionen för svenska språket vid Göteborgs universitet, MISS, 65.) Göteborg.

Josephson, Olle 2022. *Språkpolitik*. Andra upplagan. Stockholm: Morfem.

Lönnroth, Harry 2018. Portalen svenska.se – en ny digital samlingsplats för språkresurser från Svenska Akademien. *LexicoNordica* 25, 281–292.

Malmgren, Sven-Göran 2009. On production-oriented information in Swedish monolingual defining dictionaries. I: Nielsen, Sandro & Sven Tarp (eds.), *Lexicography in the 21st Century. In honour of Henning Bergenholtz*. Amsterdam/Philadelphia: John Benjamins, 93–102.

Malmgren, Sven-Göran 2014. Svenska Akademiens ordlista genom 140 år: mot fjortonde upplagan. *LexicoNordica* 21, 81–98.

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

SAG = Teleman, Ulf, Staffan Hellberg & Erik Andersson 1999. *Svenska*

Akademiens grammatik. Stockholm: Norstedts Ordbok.
SAOB = *Svenska Akademiens ordbok* 1898–. Lund: Gleerups. I: <www.saob.se> och <www.svenska.se>. Hämtad september 2022.
SAOL = *Svenska Akademiens ordlista* 2015. Fjortonde upplagan. I: <www.svenska.se>. Hämtad september 2022.
SO = *Svensk ordbok utgiven av Svenska Akademien* 2021. I: <www.svenska.se>. Hämtad september 2022.
SOB = *Svensk ordbok* 1986. Stockholm: Esselte.
Svenska.se = Svenska Akademiens ordboksportal. I: <www.svenska.se>. Hämtad september 2022.
Svensén, Bo 2004. *Handbok i lexikografi. Ordböcker och ordboksarbete i teori och praktik*. Andra upplagan. Stockholm: Norstedts.

Avledningar och sammansättningar på Synonymer.se

Viktoria Strandberg

The commercial Swedish dictionary Synonymer.se is currently one of the most frequently used digital dictionaries in Sweden and aspires to be the platform of choice for the public when it comes to questions about Swedish semantics, grammar and phonetics. This vision would necessarily demand a well-developed and accessible metalanguage, as well as a user-friendly interface. In this article, I attempt to demonstrate with which consequences the vision of the dictionary was implemented in a project on compounds and derivatives. Within this project, three new sections for compounded and derived lemmas were created in the structure of the dictionary entries: First, LEMMA *in compounds*, which presents all available compounds the lemma is part of. Second, LEMMA *consists of*, where the different parts of the lemma are listed. Third, *Related words to* LEMMA, a section including all lemmas derived from the same word family. The result shows that although the new sections expand the content of the articles and indeed make information on compounds and derivatives more accessible, they still pose challenges in terms of systematicity and redundancy. The user-friendly metalanguage and organization of the content may make the dictionary more comprehensible for its users but has at the same time resulted in deviations from linguistic and lexicographic categorizations. The conclusion is that Synonymer.se prioritizes ease of use over linguistic preciseness, a priority in alignment with the vision of the site.

KEYWORDS: compounding, derivation, digital dictionaries, Synonymer.se

1. Inledning

Den kommersiella, svenska ordbokssajten Synonymer.se utgör en av Sveriges största samtida ordböcker och har som ambition att vara tillgänglig för alla målgrupper. Kravet på användarvänlighet, här definierat som metaspråklig enkelhet och innehållsmässig lättillgänglighet, är därmed mycket högt för att locka så många användare som möjligt. Trots att ordboken lyckats bli förhållandevis välkänd bland Sveriges ordboksanvändare, har den endast ett fåtal gånger beskrivits i vetenskapliga sammanhang av lexikografer (Holmer & Sköldberg 2016; Karlholm 2020). Som tidigare lexikograf vid redaktionen för Synonymer.se kan jag i denna arti-

kel bidra med såväl ett inifrån- som ett utifrånperspektiv på ordbokssajten som lexikografisk produkt.

Artikeln presenterar arbetet med och resultatet av ett projekt om avledningar och sammansättningar på Synonymer.se i syfte att visa hur en av Sveriges mest använda digitala ordböcker balanserar bred användarvänlighet mot språkvetenskaplig korrekthet, eller, som Holmer & Sköldberg (2016:226) uttrycker det, ”Hur kan man tillvarata sajtens kommersiella intressen och kombinera dem med den lexikografiska kvaliteten?”. Artikeln syftar därmed till att lyfta fram en hittills förhållandevis dold lexikografisk process i den kommersiella ordboksvärlden, vilket är viktigt för att hela den lexikografiska mångfalden i Sverige ska beskrivas (jfr Sköldberg & Mattsson 2016:124). Med projektet om avledningar och sammansättningar som exempel, är den övergripande fråga artikeln söker svar på:

Vilka eftergifter måste Synonymer.se göra för att med språkvetenskaplig trovärdighet möta sina egna krav på användarvänlighet?

I följande avsnitt beskrivs först Synonymer.se. Därefter presenteras sajtens arbete med att presentera och integrera information om sammansättningar och avledningar, och sedan resultatet av detta projekt. Artikeln avslutas med en kort sammanfattning.

2. Synonymer.se

Den digitala ordboken Synonymer.se, ägd av företaget Sinovum Media, lanserades år 2004 och intar i dag på flera sätt en särställning bland Sveriges ordböcker. Till skillnad från exempelvis *Svenska Akademiens ordlista* (SAOL 14) har Synonymer.se inget normativt syfte, utan är rent deskriptiv. På sidan beskrivs ordbokens vision som

> att samla information om betydelser och användningar av alla svenska ord och uttryck i en digital tjänst som utöver sin primära sökfunktion möjliggör interaktivt lärande och diskussion om språkets användning och utveckling. Tjänsten ska utgöra förstahandsvalet för alla som har semantiska, grammatiska och fonetiska frågor om svenska språket (Synonymer.se 2023).

Synonymer.se består till största del av en synonymdatabas baserad på *Bonniers synonymordbok* (Walter 2000) samt tionde upplagan av *Bonniers svenska ordbok* (BSO 2011). Namnet till trots innehåller ordboken inte

bara synonymer; förutom modulen *Synonymer till* LEMMA[1] är de flesta av uppslagsorden även försedda med en definition i modulerna *Vad betyder* LEMMA*?* eller *Definition av* LEMMA.[2] En stor del av lemmana finns också tillgängliga att lyssna på. De eventuella böjningsformer ett lemma kan ha syns också i modulen *Hur böjs* LEMMA*?* och i många fall finns modulen *Hur används ordet* LEMMA*?* med exempelmeningar från bland annat dagstidningar. Användaren kan också föreslå egna synonymer, antonymer eller lemman, som granskas av en redaktör innan de läggs ut till omröstning på sajten där andra användare kan uttrycka sin åsikt om förslagen. Det slutgiltiga avgörandet ligger dock alltid hos redaktören, på samma sätt som på den användargenererade ordbokssajten Folkmun.se (Sköldberg & Wenner 2018:238). Synonymer.se är på så sätt delvis användargenererad, men i relativt liten utsträckning. Granskningsprocessen leder möjligen till att Synonymer.se inte uppdateras lika ofta som helt användargenererade ordböcker, men sajten gör betydligt fler revideringar än traditionella ordboksredaktioner. År 2022 har exempelvis information om etymologi (modulerna LEMMA *i ordbok från 1870* och *Historik för* LEMMA) samt relaterade namn och namnsdagar (modulerna LEMMA - *namnsdag & betydelse* och *Dagens namnsdag* DAGENS DATUM) lagts till. Sajten har också lanserat ett nytt format för sina kviss och korsord samt lagt till ett antal så kallade ordspel.[3]

Som Karlholm (2020:213) påpekar är Synonymer.se lättillgänglig. Den som googlat på ett någorlunda ovanligt ord har gissningsvis fått upp en länk till Synonymer.se i träfflistan; sökmotoroptimering är en stor del i sajtens framgångar (Holmer & Sköldberg 2016:220). Även internt är nästan allt innehåll i ordboksartiklarna länkat till andra, relevanta ordboksartiklar. Exempelvis är alla synonymer som anges till ett lemma klickbara. Från modulen *Vad betyder* LEMMA*?* kan användaren också länkas in till ordbokens forum. Här kan registrerade användare diskutera ordbokens

1 *Lemma* används här i betydelsen 'uppslagsord'. Synonymer.se följer inte lemma-lexem-modellen och delar följaktligen inte upp lemman med olika böjningsformer (jfr Svensén 2004:118). I stället visas alla betydelser och böjningsformer i samma artikel, se t.ex. artikeln för *bok*.

2 Skillnaden mellan modulerna *Vad betyder lemma?* och *Definition av lemma* är att den förra visar betydelsebeskrivningar från *Bonniers svenska ordbok*, medan den senare innehåller betydelsebeskrivningar skapade av redaktörer på Synonymer.se. De båda modulerna visas inte samtidigt på sidan.

3 Uppdatering gjord den 7 september 2022.

innehåll, men även andra språkrelaterade frågor. Forumet kan ses som ett led i arbetet för att minska avståndet mellan redaktion och användare (jfr Holmer & Sköldberg 2016:226) och erbjuder därmed en annan interaktion med och mellan användare jämfört med exempelvis ordboksportalen Svenska.se. Synonymer.se skulle därför per definition kunna sägas tillhöra kategorin sociala medier (jfr Weibull & Wadbring 2020:187).[4]

Förutom interaktiva funktioner som forum, korsord och kvissar kan Synonymer.se kortfattat alltså sägas bestå av olika lexikografiska databaser utarbetade vid utomstående förlagsredaktioner kompletterade med eget, lexikografiskt material. I det följande redogörs för ett projekt där sådant material skapades.

3. Projektet *Avledningar och sammansättningar på Synonymer.se*

För att skapa fler ingångar till relevanta uppslagsord för användaren, genomfördes februari–juni 2019 ett projekt i syfte att tydligare synliggöra avledningar och sammansättningar relaterade till lemmat. Sådan information fanns tidigare i undantagsfall till vissa lemman i form av språkprov, men med detta projekt ville redaktionen lyfta ut avledningar och sammansättningar i egna moduler. Målet var att användaren primärt skulle konsultera de nya modulerna för egen produktion, eftersom majoriteten av användarna vänder sig till Synonymer.se för just produktion snarare än reception (Holmer & Sköldberg 2016:222). Modulerna kan dock stötta användares reception genom att synliggöra ordled i sammansättningar och skapa sammanhang mellan avledningar tillhörande samma ordfamilj. Projektet syftade således inte till att dokumentera etymologi eller ordbildningsprocesser, trots att sådana ibland kom att synliggöras i resultatet (se avsnitt 4 nedan).

4 Jag definierar här *sociala medier* som ’internetplattformar där användarna kan bidra med eget innehåll och interagera med varandra’ (Herbert & Englund Hjalmarsson 2017:9). Med en sådan definition utgör alla användargenererade ordböcker en del av kategorin sociala medier. Många ordboksredaktioner närvarar i dag på och länkar till sociala medier (se t.ex. Biesaga 2015), men det har mig veterligen inte undersökts hur svenskspråkiga ordbokssajter själva kan fungera som sociala medier och vilken påverkan det skulle kunna ha på den lexikografiska produkten. Hur ordbokssajter kan fungera som sociala medier står inte i fokus i denna artikel, men detta är en fråga värd att överväga i framtida forskning.

I projektet deltog, förutom jag själv, ytterligare en lexikograf vid redaktionen. I ett första steg gick vi igenom varsin halva av ordbokens sammansättningar och särskrev dessa i två eller tre lemman. På så sätt kunde vi generera både de lemman en sammansättning består av och de sammansättningar ett lemma ingår i (se avsnitt 4.1 respektive 4.2 nedan). Eftersom Synonymer.se innehöll förhållandevis få affix, var den redaktionella riktlinjen att undvika att dela upp ett lemma på affixnivå såvida affixet inte redan fanns i ordboken. Detsamma gällde led i sammansättningar som ordboken där och då saknade. Vissa undantag gjordes för etablerade ord som ännu inte lagts till. En konsekvens av detta är att vissa lemman, i synnerhet avledningar, har delats upp på affixnivå (se till exempel *skönhet*[5], som delats upp i *skön* och *-het*) medan andra inte har det (exempelvis *klumpig*, som inte delats upp i *klump* och *-ig*).

I projektets andra steg grupperades de lemman i ordboken som kan sägas utgöra avledningar av en gemensam grundform (se avsnitt 4.3 nedan). Målet var att skapa mindre grupper om 5–15 lemman, för att ge användaren överskådlighet. Den övergripande redaktionella riktlinjen var användarvänlighet; användaren skulle tydligt kunna se det formmässiga sambandet mellan avledningarna och på sikt även det betydelsemässiga. Redaktörernas introspektion utgjorde grunden för indelningen, och kompletterades i vissa fall med undersökningar av de aktuella lemmanas betydelsebeskrivningar. I ett sista steg namngavs ordbokens tre nya moduler av sajtens ägare. De nya modulerna publicerades på sajten hösten 2021.

4. Resultat

Arbetet med avledningar och sammansättningar resulterade i tre nya moduler. Lemman som ingår i sammansättningar placerades i modulen LEMMA *i sammansättningar*. Om lemmat i sig är en sammansättning, synliggjordes leden i modulen LEMMA *är sammansatt av*. Modulen med avledningar relaterade till lemmat benämndes *Liknande ord till* LEMMA. De nya modulerna har, liksom sajten över lag, således ett förhållandevis enkelt, användarvänligt metaspråk och undviker facktermer som kan

5 Se dock *skönhetstävling*, som delats upp i *skönhet* och *tävling*.

vara okända för användaren, exempelvis *avledning*, *simplex* eller *ordled*.[6] Att facktermer undviks gör dessutom att Synonymer.se inte strikt behöver följa språkvetenskapliga definitioner av till exempel avledningar, utan kan i stället presentera innehållet på det sätt man anser vara mest användarvänligt. Detta till trots uppstår ändå lexikografiska utmaningar där användarvänligheten krockar med den språkvetenskapliga korrektheten. I avsnitt 4.1–4.3 redogörs för utmaningar jag själv noterade i den loggbok jag förde under projektets gång. I arbetet med denna artikel har jag dock inte kontrollerat om jag själv redigerat de moduler som jag exemplifierar med då det inte är av relevans för artikelns frågeställning.

De tre modulerna förekommer inte ofta samtidigt i en artikel, eftersom LEMMA *i sammansättningar* och LEMMA *är sammansatt av* beskriver just sammansättningar, medan *Liknande ord till* LEMMA beskriver avledningar. Som vi kommer att se är dock indelningen i sammansättningar och avledningar inte helt konsekvent. Om ett lemma behandlats som sammansättning, syns modulen LEMMA *är sammansatt av* direkt under de inledande modulerna *Synonymer till* LEMMA och *Vad betyder* LEMMA? Artiklar vars lemman hanterats som simplex har på samma placering modulerna *Liknande ord till* LEMMA och LEMMA *är sammansatt av*, förutsatt att relevant innehåll finns.

4.1 Modulen LEMMA *i sammansättningar*

Som nämndes i avsnitt 3 ovan, delades ett lemma upp i två eller tre delar förutsatt att delarna redan fanns med i ordboken eller med lätthet skulle kunna läggas till. Prefixet *miss-* fanns redan med i tionde upplagan av BSO, och *missförstå* räknades därför som en sammansättning och inte en avledning, se figur 1 nedan.

FIGUR 1. Modulen FÖRSTÅ *i sammansättningar.*

6 Som Karlholm (2020:216) påpekar skrivs dock vissa metalingvistiska kommentarer i förkortad form, t.ex. *vard.* 'vardagligt' i modulen *Synonymer till* LEMMA, vilket gör metaspråket mindre tillgängligt.

Vissa lemman, som *miss-*, ingår i många sammansättningar, vilket medför en risk för *information overload* (se t.ex. Gouws & Tarp 2017) om samtliga lemman som ordet ingår i listas i modulen. Det är rimligen av den anledningen som lemmat *miss-* och andra affixlemman saknar modulen LEMMA *i sammansättningar*. Ett vanligt förekommande ordled som däremot har ett stort antal sammansättningar listade är *sommar*, som ingår i 134 sammansättningar. För lemman som dessa visas därför endast tio sammansättningar åt gången och användaren kan få se ytterligare tio genom att klicka på *Visa fler*. Man kan fråga sig om användaren är intresserad av att klicka sig igenom så många sammansättningar, eller om modulen i fall som dessa är mer till hjälp för den lexikografiskt eller lexikologiskt inriktade forskaren. Oavsett vilket är det mig veterligen unikt för Synonymer.se att ange samtliga sammansättningar ett lemma ingår i.

4.2 Modulen LEMMA *är sammansatt av*

När sammansättningarna delades upp fanns, som tidigare nämnts, inget krav på att de skulle delas upp i endast två delar. Lemmat *förståsigpåare* delades därför upp i de tre lemmana *förstå*, *sig* och *på*, se figur 2 nedan. Suffixet *-are* fick inte plats i uppdelningen trots att det finns med som eget lemma i ordboken, då orden kunde delas upp i max tre delar. På samma sätt fick *bergochdalbana* delas upp i *berg* + *och* + *dalbana* för att passa in i formatet.

FIGUR 2. Modulen FÖRSTÅSIGPÅARE *är sammansatt av.*

Som nämndes ovan delades vissa ord upp även på affixnivå, eftersom dessa affix redan fanns i ordboken. Lemman innehållande affix som *-het* och *-aktig* kunde därför delas upp i dessa led, medan affix som *-ig* och *im-* inte skrevs ut. Inte heller fogemorfem i ord som *barnavårdscentral* och *begreppsordbok* synliggjordes. Redaktörerna har också gjort olika bedömningar och inte skrivit ut alla existerande affix även där detta varit möjligt. Till exenpel är *bekymmersam* uppdelat i *bekymmer* och *-sam*,

medan *missk*ötsam är strikt binärt indelad i *miss-* och *skötsam*, trots att *-sam* är ett i ordboken tillgängligt affix. För att få innehållet i ordboken mer konsekvent kan man därför vid en revidering överväga att alltid skriva ut alla i ordboken existerande ordled (även om det innebär att fler än tre led behöver tillåtas) eller alltid redogöra för en binär eller möjligtvis tertiär ordbildningsprocess.

I vissa fall utgjordes ordleden av böjningsformer av ordbokens uppslagsord, exempelvis *bosatt*. Här vore det möjligt att dela upp lemmat i *bo* och *satt*, eftersom en av betydelserna till det polysema *satt* hänvisar till bland annat verbet *sätta*. I projektet specificerades dock inte i modulen LEMMA *är sammansatt av* vilken underbetydelse av sammansättningsdelarna som används. Användaren som söker på *bokträd* ser till exempel att ordet är sammansatt av *bok* och *träd*, men kommer inte till den specifika underbetydelse av *bok* som avses utan måste själv lista ut vilken typ av *bok* det rör sig om. Här finns redan en diskrepans mellan BSO och *Bonniers synonymordbok*, eftersom dessa inte redogör för eller numrerar polysema ord på samma sätt. Det är gissningsvis av den anledningen som redaktören bakom artikeln *bosatt* i stället delat upp ordet i *bo* och *sätta*.

Modulens innehåll hamnar ibland i gränslandet mellan ordledsinformation och etymologisk information. Lemmat *bokstav* är etymologiskt korrekt indelat i *bok* och *stav*, vilket bekräftas i modulen *Historik för* BOKSTAV. I fallet *bokstav* kan på så sätt modulen BOKSTAV *är sammansatt av* anses vara överflödig, eftersom samma information redovisas på ett utförligare sätt längre ner i artikeln. Mindre etymologiskt korrekt är uppdelningar av ord som *bortskämd* i *bort* + *skämd*. Artikeln *skämd* saknar nämligen en referens till *skämma bort* och användaren som klickar på *skämd* i modulen *bortskämd är sammansatt av* ser endast betydelsen ’oätlig, rutten’.

Vissa lånord utgör en utmaning i sammanhanget. En del kan överhuvudtaget inte delas upp, som *avhängig* som felaktigt delats upp i *av* och *hängig*. Å andra sidan kunde *baglady* delas upp i *bag* + *lady*, eftersom båda dessa redan fanns med som separata lånord, medan *backslash* inte indelades i två ordled. En sådan uppdelning vore dock möjlig då *slash* redan existerar som uppslagsord, och *back* finns i de engelska ordböcker som Synonymer.se tillhandahåller. I nuläget finns inga länkningar mellan svenska och engelska på sajten, men möjligheten finns för såväl modulen LEMMA *är sammansatt av* som för *Historik för* LEMMA.

4.3 Modulen *Liknande ord till* LEMMA

Lemman som alla är avledningar av en gemensam grundform placerades i modulen *Liknande ord till* LEMMA. Figur 3 nedan visar de ord som angetts som avledningar till *förstå*, till exempel *förståelig* och *förståelighet*. Den som söker på *förståelig*, får upp samma modul som till *förstå* men med *förståelig* i modulnamnet och *förstå* inuti själva modulen.

FIGUR 3. Modulen *Liknande ord till* FÖRSTÅ.

Bortsett från *förståelsefull* är samtliga lemman i figur 3 avledningar. Redaktören har här valt att även ta med avledningar prefigerade med *o-*, men utelämnat avledningar med *miss-*; dessa har i stället fått utgöra en egen avledningsgrupp, trots att den gruppen är betydligt mindre än de avledningar som prefigerats med *o-*. I avsnitt 4.1 ovan såg vi att avledningen *missförstå* i stället räknas upp som en sammansättning under *förstå*. Prefixet *miss-* har alltså behandlats som ett simplex av redaktören. En sådan inkonsekvens skulle kunna göra det svårt för användaren att se skillnaden mellan modulerna *Liknande ord till* FÖRSTÅ och FÖRSTÅ *i sammansättningar*.

Den redaktör som arbetar med modulen *Liknande ord till* LEMMA behöver även bestämma hur stor hänsyn som ska tas till etymologi. Exempelvis är *klarera* en avledning av det tyska/nederländska *klar* (SAOB), men de båda orden har ett betydelsemässigt hierarkiskt förhållande där *klarera* innebär att göra ett fartyg klart för in-/avsegling. Det är gissningsvis därför som *klar* och *klarera* inte har grupperats tillsammans.

5. Summering och framåtblick

De flesta jag frågat har i något sammanhang använt Synonymer.se, i synnerhet yngre personer. Ordboken är utan tvivel en av Sveriges mest använda, förmodligen den mest använda digitala ordboken (jfr Holmer &

Sköldberg 2016:218), men relativt lite uppmärksammad i vetenskapliga sammanhang. Synonymer.se har på relativt kort tid skapat en förhållandevis känd, lättillgänglig och metaspråkligt enkel ordbokssajt där innehåll ofta uppdateras och läggs till. Det har dock skett med vissa lexikografiska eftergifter. I detta bidrag har jag tagit upp några av dessa.

Avsaknaden av på förhand bestämda tydliga redaktionella riktlinjer har resulterat i att lemman grupperats och delats upp på olika sätt under arbetet med modulerna om sammansättningar och avledningar. Det saknas exempelvis en systematik i modulen LEMMA *är sammansatt av*; vissa lemman är uppdelade även på affixnivå, medan andra enbart är uppdelade i självständiga ord. Modulens innehåll överlappar även ibland med de etymologiskt inriktade modulerna. I modulen LEMMA *i sammansättningar* redogörs mycket utförligt för de sammansättningar ordet ingår i, men man kan fråga sig hur relevant och överskådlig modulen är för användare utan specialkunskaper. Bristen på systematik i de nya modulerna upptäcks möjligen endast av mer avancerade ordboksanvändare, men skulle kunna göra att ordbokens innehåll ses som det Sköldberg & Mattsson (2016) kallar *fullexikografi*.

Synonymer.se tilldelar artiklarna ett ganska stort utrymme fördelat på flera olika moduler. Jämfört med senaste upplagan av *Svensk ordbok utgiven av Svenska Akademien* (SO, 2021), som har en post per artikel om sammansättningar och avledningar, visar Synonymer.se liknande information i tre olika moduler. För att nå de nya modulerna, som placerats längre ner på sidan, tvingas användaren skrolla en del. Man kan därför fråga sig hur användaren upptäcker de nya modulerna, och om de totalt sett många modulerna per artikel skapar en alltför tät mikrostruktur och bidrar till information overload. Relevansen för användarna skulle kunna studeras i användarundersökningar i form av enkäter och loggfilsanalyser. Hur upptäcks och används egentligen de nya modulerna? Och hur kan innehållet i dem bli pedagogiskt tydligt för användaren? Modulen *Vad betyder* LEMMA*?* innehåller en informationssymbol som användaren kan klicka på. Den ruta som kommer upp ger information om vad innehållet är baserat på och hur användaren kan förstå strukturen i modulen. Liknande information skulle med fördel kunna läggas in i inte bara de nya modulerna, utan samtliga moduler på Synonymer.se, för att möta sajtens höga ambitioner om användarvänlighet.

Litteratur

Biesaga, Monika 2015. What can a social network profile be used for in monolingual lexicography? Examples, strategies, desiderata. I: Kosem, Iztok, Miloš Jakubiček, Jelena Kallas & Simon Krek (red.), *Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of the eLex 2015 conference, 11-13 August 2015, Herstmonceux Castle, United Kingdom*. Ljubljana: Trojina, Institute for Applied Slovene Studies, 105–122.

BSO = Sjögren, Peter A. & Irene Györki 2011. *Bonniers svenska ordbok*. 10 uppl. Stockholm: Bonnier Fakta.

Folkmun.se. <http://folkmun.se>. Hämtat februari 2023.

Gouws, Rufus H. & Sven Tarp 2017. Information overload and data overload in lexicography. *International Journal of Lexicography* 30:4, 389–415.

Herbert, Ingrid & Helena Englund Hjalmarsson 2017. *Skriva i sociala medier. Handbok för digitala redaktörer*. Stockholm: Prodicta.

Holmer, Louise & Emma Sköldberg 2016. Synonymer.se i fokus. I: Gustafsson, Anna W., Lisa Holm, Katarina Lundin, Henrik Rahm & Mechtild Tronnier (red.), *Svenskans beskrivning 34*. Lund: Lund University Press, 215–228.

Karlholm, Annika 2020. Synonymer.se – Sveriges mest användbara ordbok? *LexicoNordica* 27, 213–228.

SAOB = *Ordbok över svenska språket, utgiven av Svenska Akademien*. 1898–. I: <saob.se>. Hämtat februari 2023.

SAOL 14 = *Svenska Akademiens ordlista över svenska språket*. (Upplaga 14, 2015). I: <svenska.se>. Hämtat februari 2023.

Sköldberg, Emma & Christian Mattsson 2016. Ful- och finlexikografi? Om ordboksverksamhet i Sverige i dag och i morgon. I: Hannesdóttir, Anna (red.), *Framtidens lexikografi. Rapport från ett symposium i Göteborg 5 oktober 2012*. (Meijerbergs arkiv för svensk ordforskning 42.) Göteborg: Meijerbergs institut för svensk etymologisk forskning, 111–129.

Sköldberg, Emma & Lena Wenner 2018. Amatörlexikografiska insatser på sajten *Folkmun.se*. I: Svavarsdóttir, Ásta, Halldóra Jónsdóttir, Helga Hilmisdóttir & Þórdis Úlfarsdóttir (red.), *Nordiske Studier i Leksikografi* 14. Reykjavik: Nordisk Forening for Leksikografi, 237–245.

SO = *Svensk ordbok utgiven av Svenska Akademien*. 2021. 2 uppl. I: <svenska.se>. Hämtat februari 2023.

Svensén, Bo 2004 [1987]. *Handbok i lexikografi. Ordböcker och ordboksarbete i teori och praktik*. 2 uppl. Stockholm: Norstedts Akademiska Förlag.

Svenska Akademiens ordböcker. <http://svenska.se>. Hämtat februari 2023.

Synonymer.se. <http://synonymer.se>. Hämtat februari 2023.

Walter, Göran 2000. *Bonniers synonymordbok*. 3 uppl. Stockholm: Bonnier.

Weibull, Lennart & Ingela Wadbring 2020. *Det svenska medielandskapet. Traditionella och sociala medier i samspel och konkurrens*. Stockholm: Liber.

Lexikon för ett skriftlöst språk

Jan-Olof Svantesson

In this article I discuss problems related to the compilation of dictionaries for understudied languages that do not have a generally accepted written form. This task is in most cases undertaken by linguists with the purpose of describing and analyzing a language or by missionaries whose ultimate goal is to translate the Bible or other holy scriptures. As a case study, the structure of two different versions of a dictionary for Kammu, a minority language in Northern Laos, is described. One dictionary is primarily intended for linguists and other researchers, and the other one for the speakers of the language.

NYCKELORD: skriftlösa språk, minoritetsspråk, kammu

1. Syfte

Artikelns syfte är visa på skillnader och likheter mellan lexikon för språk som saknar ett etablerat skriftspråk och för språk som har en allmänt använd skrift genom att beskriva arbetet med att utarbeta lexikon för kammu, ett minoritetsspråk i norra Laos. Bakgrunden är att jag ingår i en grupp vid Lunds universitet som forskar om kammufolket och som gruppens lingvist har jag bland annat arbetat med lexikonen.

Några forskare som själva har arbetat med lexikon för outforskade språk har satt in sitt arbete i ett lingvistiskt och lexikografiskt sammanhang, t.ex. Connell (1998) och Mosel (2004). Som Connell (s. 231) påpekar behandlas lexikografin mycket styvmoderligt i allmänlingvistiska handböcker och läroböcker. Omvänt behandlar den lexikografiska litteraturen sällan de speciella problemen med lexikon för outforskade språk. Bo Svenséns *Handbok i lexikografi* (2004) innehåller ingenting om detta, och inte heller *LexicoNordica* eller *Nordiske studier i leksikografi*, kanske med undantag för några artiklar och recensioner av dialektlexikon (t.ex. Wendt 2007), som skulle kunna räknas till denna kategori.

2. Kammuspråket

Kammu tillhör den austroasiatiska språkfamiljen och talas av över 500 000 personer, främst i norra Laos. Det finns ingen allmänt använd skrift för språket, och det förekommer ingen skolundervisning på kammu i Laos där i princip alla skolor undervisar på majoritetsspråket laotiska.

Kammuprojektet vid Lunds universitet startades i början av 1970-talet av Kristina Lindell som i första hand arbetade med muntlig litteratur. Till projektet knöts den infödde kammutalaren Kàm Ràw (Damrong Tayanin), som flyttade till Lund och blev permanent medarbetare i projektet. Flera andra forskare har medverkat i projektet; se Lundström & Svantesson (2005).

3. Skriftlösa språk

Vad är ett skriftlöst språk? Enligt *Ethnologue*, en förteckning över världens språk som ges ut av den världsomspännande amerikanska missionärsorganisationen Summer Institute of Linguistics (SIL) har ca 4 000 språk, långt mer än hälften av världens cirka 6 000 språk, ett skriftspråk. Det är säkert korrekt, men ger en skev bild av sakernas tillstånd, eftersom det i de flesta fall rör sig om skriftspråk som används eller har använts endast i begränsad omfattning. För att få en bättre bild kan man vända sig till handböcker om identifiering av språk för bibliotekarier, t.ex. Giljarevskij & Grivnin (1964). De redovisar typiskt 200–300 språk, och en realistisk gissning är att antalet skriftspråk som är allmänt spridda bland talarna och som är i dagligt bruk i tidningar och böcker, i skolundervisning och i andra officiella sammanhang bara är drygt 200. Med ”skriftlösa språk” menar jag i fortsättningen språk som inte har ett allmänt använt skriftspråk. Några missionärer och lingvister (inklusive vår grupp i Lund) har givit ut mindre skrifter på kammu, men de har haft begränsad spridning och det har inte förekommit någon standardisering mellan de olika skriftsystem som har använts, så kammu är ett skriftlöst språk i denna bemärkelse.

4. Vem utarbetar lexikon för skriftlösa språk?

Det är nästan bara lingvister och missionärer som utarbetar lexikon för skriftlösa språk. Ett exempel på en framgångsrik missionärslingvist är svensk-amerikanen Ola Hanson (1864–1929), född i Åhus, som 1895

skapade en skrift för kachin, ett minoritetsspråk i Burma (Myanmar). Som förberedelse för att översätta Bibeln skrev han ett 750-sidigt kachin-engelskt lexikon som gavs ut 1906 och fortfarande kommer ut i nya upplagor.

5. Målgrupper

Man kan säga att lexikon över outforskade språk till en början har lexikonförfattaren själv som målgrupp. För att kunna analysera eller lära sig ett språk måste man samla in ett ordmaterial och göra upp en ordlista i någon form. Några lexikonprojekt kommer inte längre än så. Ett exempel är det omfattande (ca 7 000 ord) kalmuckisk-svenska lexikon som sammanställdes av Cornelius Rahmn (1785–1853) som var missionär bland kalmuckerna, ett mongoliskt folk vid Volgas nedre lopp, under åren 1819–23. Han tvingades sedan lämna Ryssland och återvände så småningom till Sverige som kyrkoherde i Kalv i Västergötland. Lexikonet, som finns bevarat som manuskript i Uppsala universitetsbibliotek, gavs långt senare ut i redigerad form som Rahmn (2012). Det är det tidigaste större lexikonet över kalmuckiska och ett språkhistoriskt värdefullt dokument.

Målgruppen av personer som, liksom lexikonförfattaren, vill lära sig språket ifråga är oftast ganska liten eftersom det handlar om minoritetsspråk som saknar skriftlig litteratur. Det kan vara missionärer som vill sprida sin religion, men också forskare som arbetar med folkets kultur och som gör upp egna ordlistor över ”sina” språk. Många lexikon som ges ut av språkvetare har, liksom numera Rahmns lexikon, andra språkvetare som målgrupp, i första hand sådana som arbetar med jämförande historisk språkforskning men också inom områden som språktypologi eller semantik.

En helt annan målgrupp är talarna av språket ifråga. Ett skriftlöst språk är så gott som alltid ett minoritetsspråk och talarna kan tänkas behöva ett lexikon som talar om vad ett ord på deras språk heter på majoritetsspråket. Men talare av skriftösa språk är ofta minst tvåspråkiga, har gått i skola på majoritetsspråket och behärskar i många fall majoritetsspråket lika bra som – eller bättre än – sitt eget språk. Det är ofta stigmatiserat att tala ett minoritetsspråk och många talare av skriftlösa språk använder majoritetsspråket oftare än sitt modersmål. Det kan leda till att språket inte lärs ut till barnen, får färre och färre talare, och så småningom dör ut. I sådana fall kan ett lexikon bidraga till en revitalisering av språket

eftersom det visar att det är möjligt att skriva språket och tillhandahåller en användbar standard för skriftspråket. Ett exempel är det omfattande revitaliseringsarbete som görs av Suwilai Premsrirat och hennes medarbetare vid Mahidol-universitetet utanför Bangkok. Genom att utarbeta lexikon och läromaterial har de lyckats få in flera mindre minoritetsspråk i skolundervisningen och kunnat vända en nedåtgående trend i användningen av några av dem (Premsrirat & Hirsh 2018).

Ett lexikon för ett skriftlöst språk kan alltså vara en viktig symbol som visar att språket inte bara är ett talspråk med låg status, utan ett språk som faktiskt kan skrivas – vilket för många talare av minoritetsspråk inte är en självklarhet. I bästa fall kan ett tidigare diskriminerat språk och folk få en helt ny status. Ett exempel är Ola Hansons kachinlexikon som tillsammans med hans grammatik och bibelöversättning ledde till att kachinerna, som så gott som alla var analfabeter när han kom till Burma, nu är allmänt läskunniga på sitt eget språk (Sword 1954).

För kammu har vi försökt lösa motsättningen mellan målgrupperna genom att göra två olika versioner av lexikonet, ett från kammu till engelska som i första hand riktar sig till lingvister och andra forskare, och ett från kammu till laotiska för modersmålstalarna. Mosel (2004:41) nämner denna möjlighet men ger inga exempel på att den har använts. I avsnitt 6 beskrivs det kammu-engelska lexikonets struktur och i avsnitt 7 det kammu-laotiska.

6. Det kammu-engelska lexikonet

I detta avsnitt beskrivs det kammu-engelska lexikon som har utarbetats i Lund (Svantesson et al. 2014) som ett exempel på lexikon för skriftlösa språk skrivna av (och för) lingvister och andra forskare.

För att överhuvudtaget kunna skriva ner ord måste man analysera ljudsystemet och skapa en skrift som återger språket på ett adekvat sätt. En inflytelserik artikel om hur man skapar nya skriftspråk är Smalley (1964), som ger fem ”maximer” om hur ett välfungerande skriftspråk ska konstrueras. Här tar jag inte upp detta problem utan koncentrerar mig på lexikografin.

Kammuprojektet började med att Kristina Lindell spelade in folksagor som berättades av olika kammutalare som befann sig i norra Thailand men ursprungligen kom från Laos. Tillsammans med Kàm Ràw, som själv

var en framstående sagoberättare, transkriberade hon inspelningarna. De skrevs ned med det latinska alfabetet med tillägg av en del IPA-tecken och samma skrivsätt används i lexikonet (figur 1).

♦**Hɔ́ɔ Yàŋ** [Lao *yaŋ*] a Tibeto-Burman people in Laos

hɔ́ɔ² [Lao *hɔɔ*, Lü *hɔ*] ‡ building [shaman song]

hɔ́ɔ³ oh!

hɔ́ɔc [Red. *híc*] to whittle, to cut with the knife blade held almost parallel to the cut surface ●~ *kháal* cut bamboo strips as warp for weaving baskets ●~ *plɔ̀ɔŋ* cut rattan ●~ *sʔɔ́ɔŋ* whittle wood

►CAUS **pɛ̀'hɔ́ɔc** to make sb whittle

►NOMp **rc'hɔ́ɔc** [Red. *rc'héc*] action of whittling

►NOMi **hrnɔ̀ɔc 1** action of whittling **2** things made by whittling e.g. bamboo strips

hɔ́ɔk¹ [Lao *rɔɔk*, Lü *rɔk*] ‡ squirrel [*krùu*] =Km *prɔ́ɔk*

hɔ́ɔk² [Lao *hɔɔk*, Lü *hrɔk*^{L1}] ‡ sword [*krùu*] =Km *kmáŋ*

hɔ́ɔl srmà

hɔ́ɔl to smear, to paint ●*kàaŋ* ~ painted house ●~ *spí* paint red marks on one's face >*spí*

♦**hɔ́ɔl lɔ̀** ceremony for renewing the strength during a buffalo sacrifice >*tráak*

♦**hɔ́ɔl srmà** paint a sign with lime (*púun*) on the back of a person with fever, in order to cure him

hɔ́ɔm¹ [Lü *rɔm*] to tie, to tie up; to gather ●~ *klə̀* put up one's hair to a topknot ●*klə̀* ~ topknot ◊ ●~ *trɔ̀* bind torches

♦**hɔ́ɔm kmɔ̀póŋ** tie a charm to a bamboo strip for curing headache >*cú kmpóŋ*

►CAUS **pɔ̀nhɔ́ɔm** to make sb tie up e.g. hair

►NOMp **rmhɔ́ɔm** [Red. *rmhéem*] action of tying

►NOMi **hrnɔ̀ɔm 1** [=, *tí*] bundle ●*hʔé mòoy* ~ one bundle of firewood **2** [*lém*] bamboo strip for tying things ◊*prìiŋ, syɔ́ɔc* ●*hɔ́ɔm yʌ̀ʌ* ~ tie with a bamboo strip ●~ *mòoy tí* one bundle of bamboo strips

♦**hrnɔ̀ɔm klə̀** hairknot

hɔ́ɔm² [Lao *hɔɔm*, Lü *hɔm*] aromatic herb, onion (*phák* ~)

♦**hɔ́ɔm búa** "lotus onion" Japanese bunching onion, *Allium fistulosum* [Botanical garden]

Traditionally not used by the Kammu.

♦**hɔ́ɔm déɛŋ** "red onion" red onion

Traditionally, the Kammu do not use onions on their food, except on meat salad, so they do not grow much onion.

♦**hɔ́ɔm dían** field mint, *Mentha arvensis* L. (Lamiaceae)

Mint which is used on meat salad and chicken chilli sauce.

♦**hɔ́ɔm kháaw** "white onion" garlic, *Allium sativum* L. (Alliaceae)

Traditionally, the Kammu do not eat garlic, but now some villagers grow it for sale and own use.

♦**hɔ́ɔm pɔ́ɔm-páa** [Lao *hɔɔm pɔɔm*^{2}] long coriander, *Eryngium foetidum* L. (Apiaceae)

Wild herb which grows at open places in the forest. The leaves are used as a spice on eggplant salad, on roasted fish, on taro stew (*sáŋ tàŋ cró*) with buffalo hide, and on chilli sauce with roasted fish.

hɔ́ɔñ* expressive root:

►STAT **tñhɔ́ɔñ | rñhɔ́ɔñ ↓ tñhéñ | rñhéñ** dark in one|many places (*hlŋìñ* ~)

►PCT **thɔ́ɔñ ↓ théñ** become dark

►IRR **slthúuñ-slthɔ́ɔñ ↓ slthúuñ-slthéñ** become dark many times in different places ●*prlìa pìit* ~ all the fires went out one by one

hɔ́ɔŋ¹ [Lao *hɔɔŋ*^{2}, Lü *hɔŋ*^{2}] [=] room

♦**hɔ́ɔŋ krùa** kitchen

♦**hɔ́ɔŋ nám** hall

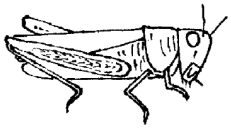
♦**hɔ́ɔŋ òm** bathroom, toilet

♦**hɔ́ɔŋ sís** sleeping room

hɔ́ɔŋ² [Lao *rɔɔŋ*^{2}, Lü *rɔŋ*^{2}] ‡ to shout, to call =Km *héet*

hɔ́ɔŋ³ R reason, ability =Y *sén*

hɔ́ɔs short-horned grasshopper, locust (subfamily Caelifera)



hɔ́ɔs kɨŋ

♦**hɔ́ɔs kɨŋ** "rotten egg grasshopper" spotted grasshopper (*Aularches miliaris*)

Brown grasshopper with spotted wings. It lives in bushes. It smells like rotten eggs and is not eaten.

♦**hɔ́ɔs kóoŋ** "frozen grasshopper" a kind of grasshopper

Large green edible locust which eats grass and rice leaves. People collect them when they are weeding the fields, roast them and let the children eat with rice.

♦**hɔ́ɔs màt pɛ̀** "goat eye grasshopper" migratory locust (*Locusta migratoria*)

Brown inedible grasshopper with large wings.

♦**hɔ́ɔs plɔ̀** a kind of grasshopper

Grey grasshopper with spotted wings. It is similar to *hɔ́ɔs túus*.

♦**hɔ́ɔs pɛ̀yàak** "sasagrass grasshopper" a kind of grasshopper

Green grasshopper which lives in the grass.

♦**hɔ́ɔs pɛ̀yɔ̀ɔŋ** "dragon grasshopper" a kind of grasshopper

Small green grasshopper which lives in the grass and on rice plants, and eats the leaves.

♦**hɔ́ɔs tɾlèŋ-téŋ** "wagtail grasshopper" a kind of grasshopper

FIGUR 1. En sida i det kammu-engelska lexikonet.

Eftersom talare av olika dialekter hade spelats in uppstod problemet hur vi skulle förhålla oss till dialektformerna. Det finns tre huvuddialekter i Laos: nordkammu, västkammu och östkammu (se Svantesson & Holmer 2015 för en allmän översikt över kammuspråket). Skillnaderna är inte särskilt stora och dialekterna är inbördes fullt begripliga. Östkammu var dåligt representerat i det inspelade materialet och de flesta av våra sagoberättare talade olika underdialekter av nordkammu, t.ex. Yùan som var Kàm Ràws modersmål. Det finns en del fonologiska skillnader mellan talarna, men vi beslutade oss för att inte ange olika uttalsvarianter utan använda Kàms uttal i lexikonet, även för ord som egentligen inte tillhör Yùan-dialekten. Detta gjordes delvis av praktiska skäl eftersom vi inte hade möjlighet att arbeta med talare av olika dialekter på ett systematiskt sätt, men också för att orden i lexikonet fonologiskt och morfologiskt skulle bilda ett konsistent system. I skriftlösa språk som kammu finns det ju inget standardspråk, men att använda en enskild talares uttal innebär indirekt en sorts standardisering (Mosel 2004:42).

6.1. Material och ordurval

Det kammu-engelska lexikonet innehåller 498 sidor och trycktes i 200 exemplar. Antalet ord är cirka 14 900, varav 3 300 är avledningar och 3 150 är sammansättningar. Materialet för lexikonet var till en början muntlig litteratur, främst folksagorna som hade spelats in och översatts till engelska. Kàm Ràw var förutom sagoberättare också en god sångare med en stor repertoar av traditionella kammusånger. Hans och andras sånger analyserades i samarbete med musiketnologen Håkan Lundström och ordförrådet i sångerna togs in i lexikonet. Ordförrådet i böner och andra rituella texter vid ceremonier i samband med jordbruksåret och vid giftermål, barnafödsel, begravningar och andra tillfällen togs också in i lexikonet. Vi gick också systematiskt igenom olika semantiska fält med Kàm Ràw, exempelvis ord som hör samman med jordbruksåret, med jakt eller med husbyggnad, liksom släktskapsord, kalenderterminologi, musikterminologi och annat; Mosel (2004:45) kallar det för ”Active Elicitation”. Kàm identifierade också ett stort antal avledda ord genom sådan självelicitering.

Kàm utbildades i sin ungdom till schaman, vilket innebar att han förutom traditionell örtmedicin fick lära sig olika rituella texter som böner

och besvärjelser som riktar sig till andar som ansågs orsaka sjukdomar. Många av dessa innehåller inga inhemska kammuord utan består helt och hållet av ord lånade från äldre stadier av laotiska och är obegripliga för vanliga kammutalare. Sådana ord anses vara farliga om de kommer i orätta händer, men Kàm accepterade att de togs in i det kammu-engelska – men inte i det kammu-laotiska – lexikonet. Kàms lärare dog innan Kàm hade invigts till schaman och han hade därför inte lovat att hålla denna kunskap hemlig (jämför Mosel 2004:47 om tabuerade ord). Dessa ord markeras i lexikonet med symbolen "‡". Ett exempel i figur 1 är *hɔ́ɔk*[2] 'svärd', som på vanlig kammu heter *kmáŋ*. Oftast anges också typen av källa, i det här fallet en besvärjelse (*krùu*).

Ett problem är förhållandet till majoritetsspråket. Liksom många talare av minoritetsspråk är de flesta kammuer minst tvåspråkiga, i detta fall på kammu och laotiska. All skolundervisning är på majoritesspråket laotiska, som de flesta kammuer numera lär sig i skolan eller tidigare. Alla ord för moderna företeelser som inte ingår i traditionell kammukultur tas över direkt från laotiska och vi bedömde det som ogörligt att ta med dem i lexikonet eftersom i stort sett alla laotiska ord kan användas i kammu. Ett problem är att det finns många gamla fullt etablerade lånord som självklart måste tas med i lexikonet, och gränsen mellan äldre och nyare lånord är ofta svår att dra.

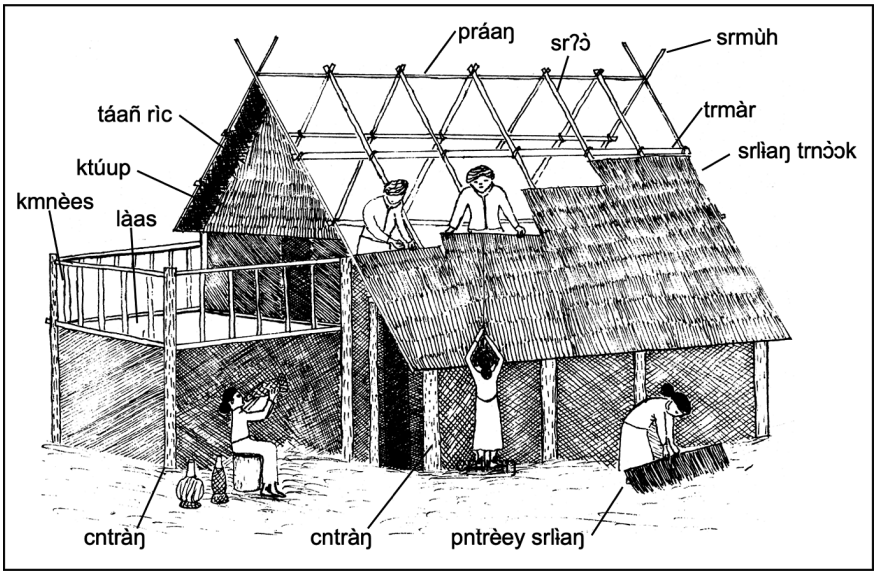
Sammanfattningsvis kan sägas att ordurvalet ändå var enkelt, alla ord som vi påträffade och som på något sätt anknöt till traditionell kammukultur togs med i lexikonet.

6.2. Ordens betydelser

Lexikonet utgick från början från folksagor där ordens betydelser ofta är ganska tydliga. Under inspelningarna ordnade Kristina Lindell seminarier med sagoberättarna där svårbegripliga passager diskuterades. I sånger, böner och rituella texter kan orden vara svårare att förstå och ha en mer kontextbunden betydelse, och vi har där fått lita på Kàm Ràws djupa kunskaper om sin kultur.

Ett problem som sannolikt har drabbat oss då och då och är svårt att göra något åt är att orden ges för snäv betydelse. Även om den betydelse som ges i lexikonet är korrekt i den kontext där ordet dök upp – t.ex. i en folksaga – är betydelseomfånget kanske större, något som ofta har visat

sig efterhand under lexikonarbetet. Vårt utgångsmaterial är begränsat, liksom den tid vi har haft till förfogande, vilket har gjort det omöjligt att göra helt uttömmande beskrivningar av ordens betydelser.



FIGUR 2. En teckning i det kammu-engelska lexikonet som visar ord för olika delar av ett hus.

Vid exempel som kommer från folksagor, sånger, böner, rituella texter och så vidare har typen av källa markerats inom hakparentes, till exempel ”[tale]”. Alla ordspråk, talesätt, gåtor och andra fasta uttryck som vi känner till ges i lexikonet, i allmänhet under det första innehållsordet. I ett appendix förtecknas geografiska namn (byar, berg, floder och annat) i Yùan-området. Lexikonet illustreras med ca 300 teckningar, till största delen utförda av Kàm Ràw. De flesta är bilder som illustrerar betydelsen hos olika ord, som i figur 2.

6.3. Avledningar och sammansättningar

Kammu är ett isolerande språk som helt saknar böjningsmorfologi. Däremot finns det en utvecklad avledningsmorfologi som inte är helt produktiv och därför måste alla avledda ord förtecknas i lexikonet. Eftersom lexikonet i första hand vänder sig till lingvister valde vi att lista avledningar till-

sammans med det ord de avleds från. Ett exempel är ordet *hɔ́ɔm*[1] 'binda' i figur 1, som kan bilda kausativ med prefixet *pn-* (*pǹhɔ́ɔm*) och nominaliseras på två olika sätt: mer abstrakt ('bindande') med prefixet *r*C-, där C är ordets finalkonsonant (*rmhɔ́ɔm*), och mer konkret ('bamburemsa') med infixet *-rn-* (*hrnɔ̀ɔm*). Alla dessa listas under lemmat *hɔ́ɔm*[1] (figur 1) där avledningsinformationen är lättillgänglig för en lingvist som dessutom får upplysningar om vilken sorts avledning det handlar om, t.ex. CAUS för kausativ och NOMi för nominalisering med ett infix. För en användare som har träffat på ordet *hrnɔ̀ɔm* och vill veta vad det betyder är proceduren däremot litet mer omständlig: Slår man upp *hrnɔ̀ɔm* på dess alfabetiska plats finns där en hänvisning till *hɔ́ɔm*. Denna organisation av lexikonet resulterar i ett stort antal hänvisningar. Sammansatta ord listas direkt efter det första ordet i sammansättningen (som i *hɔ́ɔm búa* 'piplök' under *hɔ́ɔm*[2] i figur 1).

6.4. Grammatiska upplysningar

Lexikonet innehåller en del grammatiska upplysningar. Ordklasser kan inte, som i exempelvis svenska eller engelska, definieras med utgångspunkt från böjningsmorfologin, eftersom sådan saknas i kammu, utan måste grundas på syntaxen. Eftersom vi inte hade gjort någon ingående syntaktisk undersökning av språket när lexikonet utarbetades kunde vi inte ge någon vetenskapligt grundad indelning i ordklasser utan nöjde oss med att sätta ut det engelska infinitivmärket *to* i översättningen av ord som motsvarar engelska verb och vi har också försökt ge exempel som visar verbens argumentstruktur. Se t.ex. *hɔ́ɔm*[1] 'binda' i figur 1, där exemplet *hɔ́ɔm klə̀* 'sätta upp håret' visar att verbet konstrueras med ett (omarkerat) direkt objekt (*klə̀* 'hår').

För substantiv anges vilken klassifierare som används. I kammu liksom i många andra språk i Öst- och Sydöstasien kan man inte kombinera ett substantiv direkt med ett räkneord utan man använder en klassifierare. Exempelvis anges att klassifieraren till *hrnɔ̀ɔm* 'bamburemsa' (under *hɔ́ɔm*[1] i figur 1) är *lém*. Det innebär att till exempel 'två bamburemsor' heter *hrnɔ̀ɔm pàar lém* (där *pàar* är räkneordet 'två') och inte bara **hrnɔ̀ɔm pàar*. Det finns ett tjugotal olika klassifierare. Exakt vilken klassifierare som krävs för ett visst substantiv är ofta inte självklart utan måste anges i lexikonet. I vissa fall är det självklart, t.ex. tar alla växter klassifieraren *túut* så det har inte satts ut i lexikonet.

Många ord, mest verb och adjektiv, kan reduplicera med förändrad vokal eller rim, vilket anger intensifierad betydelse, att något görs snabbt eller kraftfullt. Alla kända sådana former förtecknas i lexikonet. Exempelvis är *hɔ́ɔc-híc* en intensivform av *hɔ́ɔc* (figur 1).

6.5. Ordens ursprung

Om ett ord har påträffats bara i en annan dialekt än Yùan anges det; exempelvis anges att ordet *hɔ́ɔŋ*[3] är Rɔ̀ɔk-dialekt (markerat med ”R” i figur 1) och att motsvarande ord i Yùan-dialekten (”Y”) är *sén*.

Många ord har under tidernas lopp lånats in från majoritetsspråket laotiska, och också från lü, ett språk som talas i Yunnan-provinsen i Kina och är ganska nära släkt med laotiska. Särskilt i rituellt språk finns det ord som kommer från pali och de har angetts i den mån vi har kunnat identifiera dem. Ordens form i de långivande språken ges inom hakparentes (se t.ex. *hɔ́ɔŋ*[1] i figur 1). Däremot har vi inte gett etymologier för inhemska austroasiatiska ord.

6.6. Encyklopedisk information

Under vårt arbete med lexikonet blev det efterhand allt tydligare att den traditionella kammukulturen som är grundad på risodling i svedjor på bergssluttningarna höll på att försvinna. Många tvingades av ekonomiska skäl att flytta närmare vägar och floder och många byar avfolkades. I Kàms hemby Rmcùal utfördes jordbruksårets riter för sista gången år 1989 och ett tiotal år senare hade byn avfolkats helt (Évrard 2012). De flesta av Kàms generationskamrater hade då gått till sina förfäder och han var troligen en av de sista som hade detaljerad kunskap om den traditionella kammukulturen i området, och definitivt den ende som hade möjlighet att dokumentera den för framtiden. Detta gjorde att vi tog in hans beskrivningar av traditionellt kammuliv i lexikonet, sådant som ceremonier och riter under kammuernas livscykel och jordbruksår, beskrivningar av hur växter, djur och redskap användes och mycket annat (se t.ex. de korta beskrivningarna av olika gräshoppor under *hɔ́ɔs* i figur 1). Många besvärjelser, böner och andra rituella texter togs också in. Vi insåg att det inte var helt lyckat att publicera encyklopediska uppgifter i ett lexikon, men med de begränsade resurser vi hade var det i vilket

fall ett sätt att bevara Kàms kunskaper för framtiden. Även om beskrivningarna i första hand gällde Kàms hemby Rmcùal är vi övertygade att mycket gäller i hela Yùan-området och till och med i hela det kammutalande området.

7. Det kammu-laotiska lexikonet

I samarbete med laotiska lingvister sammanställde vi ett lexikon från kammu till laotiska som var riktat till kammutalarna (Svantesson et al. 1994). Lexikonet utarbetades med stöd från SIDA, trycktes i 1 000 exemplar och spreds till kammuer i Laos. Vi hoppades att lexikonet skulle kunna användas i skolundervisningen men våra kontakter med berörda myndigheter ledde inte till några resultat. En organisation som däremot drog nytta av lexikonet var den nationella laotiska radions redaktion för sändningar på kammu som fick ett användbart skriftspråk.

Det kammu-laotiska lexikonet baserades på vårt dåvarande ordmaterial för det kammu-engelska lexikonet, som inte gavs ut förrän 20 år senare. Det innehåller ca 10 000 ord och består av 514 sidor med betydligt mindre format än i det kammu-engelska lexikonet. Dess struktur skiljer sig på många sätt från det kammu-engelska lexikonet, beroende på de olika målgrupperna, men också eftersom den tid och arbetskraft vi kunde lägga ner var mycket mindre. Den största skillnaden är att i det kammu-laotiska lexikonet skrivs kammuorden med en dialektneutral skrift som använder det laotiska alfabetet. Trots att kammu och laotiska inte är besläktade språk är fonemförråden ganska lika, så det var inte något stort problem att skriva kammu med det laotiska alfabetet, även om uttalet av några bokstäver fick anpassas något. Uppslagsorden ges också med det latinska alfabetet eftersom Kàm hade lärt ganska många att skriva och läsa kammu på det sättet. En annan viktig skillnad är att avledningar står på sin alfabetiska plats med hänvisning från ordet de avleds från. Antalet exempel är mycket mindre och långa autentiska exempel från t.ex. folksagor saknas helt, liksom sådant som ordspråk och talesätt. Sammansatta ord är inte självständiga underartiklar utan listas som exempel. Encyklopedisk information och längre texter (riter, böner etc.) saknas helt. Det finns också mycket färre bilder.

8. Avslutning

Genom att beskriva arbetet med kammulexikonen har jag försökt visa på skillnaderna mellan lexikon för skriftlösa språk och för språk med en väletablerad skrift. Enligt min åsikt är den viktigaste skillnaden egentligen inte hur lexikonen är uppbyggda utan den symboliska betydelsen hos ett lexikon för ett skriftlöst språk och dess betydelse för språkrevitaliseringsprogram.

Till sist vill jag åter säga att det inte finns mycket kontakt mellan professionella lexikografer och ”lingvistlexikografer”. Själv är jag lexikografisk amatör, liksom de flesta jag känner till som arbetar med lexikon för skriftlösa eller föga utforskade språk. Omvänt behandlar den lexikografiska litteraturen sällan de speciella problemen med lexikon för skriftlösa eller outforskade språk. Förhoppningsvis kan mitt bidrag öka kontakterna mellan dessa grupper av lexikonförfattare.

Referenser

Connell, Bruce 1998. Lexicography, linguistics, and minority languages. *Journal of the Anthropological Society of Oxford* 29:3, 231–242.

Ethnologue. Languages of the World. 25 uppl. <www.ethnologue.com>. Hämtat juli 2022.

Évrard, Olivier 2012. Following Kàm Ràw’s trail. I: Tayanin, Damrong & Kristina Lindell (red.), *Hunting and fishing in a Kammu village. Revisiting a classic study in Southeast Asian ethnography*. Copenhagen: NIAS Press, 1–28.

Giljarevskij, Rudžero Sergeevič & Vladimir Sergeevič Grivnin 1964. *Opredelitel' jazykov mira po pis'mennostjam.* 3 uppl. Moskva: Nauka.

Hanson, Ola 1906. *A dictionary of the Kachin language*. Rangoon: American Baptist Mission Press.

Lundström, Håkan & Jan-Olof Svantesson (red.) 2005. *Kammu: om ett folk i Laos*. Lund: Lunds universitetshistoriska sällskap. (Årsbok 2006).

Mosel, Ulrike 2004. Dictionary making in endangered speech communities. *Language Documentation and Description* 2, 39–54.

Premsrirat, Suwilai & David Hirsh (red.) 2018. *Language revitalization. Insights from Thailand*. Bern: Peter Lang.

Rahmn, Cornelius 2012. *Cornelius Rahmn's Kalmuck dictionary*. Utgiven av Jan-Olof Svantesson. Wiesbaden: Harrassowitz.

Smalley, William A. 1964. How shall I write this language?. I: Smalley, William A. (red.), *Orthography studies. Articles on new writing systems*. London: The United Bible Societies, 31–59.

Svantesson, Jan-Olof & Arthur Holmer 2015. Kammu. I: Jenny, Mathias & Paul Sidwell (red.), *The handbook of Austroasiatic languages*. Leiden: Brill, 957–1002.

Svantesson, Jan-Olof, Kàm Ràw, Kristina Lindell & Håkan Lundström 2014. *Dictionary of Kammu Yùan language and culture*. Copenhagen: NIAS Press.

Svantesson, Jan-Olof, Damrong Tayanin, Kristina Lindell, Thongpheth Kingsada & Somseng Xayavong 1994. *Watcanaanukom khamu-laaw* [Kammu-laotiskt lexikon]. Lund: Lund University och Vientiane: Institute for Research on Lao Culture, Ministry of Information and Culture.

Svensén, Bo 2004. *Handbok i lexikografi. Ordböcker och ordboksarbete i teori och praktik*. 2 uppl. Stockholm: Norstedts.

Sword, Gustaf A. 1954. *Light in the jungle. Life story of Dr. Ola Hanson of Burma*. Chicago, IL: Baptist Conference Press.

Wendt, Bo-A. 2007. Lödigt lexikograferade göingemål. *LexicoNordica* 14, 341–354.

Brugernes blik på Jysk Ordbog. En brugerundersøgelse i leksikografiens tegn

Mette-Marie Møller Svendsen

Jysk Ordbog (JO) is a historical dictionary that is freely available online at jyskordbog.dk. The process of editing the extensive source material for the dictionary has spanned 90 years, and there is still plenty of work left to do for the editors of JO. The dictionary has existed online since 2000, and from 2000 until 2021 its looks and functionality have remained basically unchanged. As dictionary users' needs and expectations change over time, a user research-based study of jyskordbog.dk can help shed light on which parts of the online dictionary work well, and which parts need to be revised and updated. This article presents the background for the user research study for JO that took place in 2020–2021, the methods used for the data collection, and the results of the study. I draw connections to past lexicographical user research insights, and I discuss and compare the results concerning layout and functions to other Scandinavian online dictionaries. Suggestions are also presented for how JO can implement useful changes to its online format provided by the insights from the study.

KEYWORDS: user experience research, lexicography, purposive sampling

1. Indledning

> Olika typer av ordböcker brukas av olika typer av användare med hjälp av olika typer av strategier i samband med olika typer av språkliga aktiviteter för att söka olika typer av information som behövs i olika typer av situationer. Detta konstaterande har inte alltid varit lika självklart som det är idag: förr utformades ordböckerna ofta efter ordboksutgivarnas egna föreställningar om användarnas kompetens och behov och utan stöd av empiriska undersökningar av hur ordböcker användes (Svensén 2004:533).

Som ovenstående citat viser, så er det absolut ikke uinteressant for en ordbog at kende sine brugere – én løsning passer ikke til alle. Siden Wiegand kaldte ordbogsbrugeren 'den kendte ubekendte' tilbage i 1977 (citeret

efter Svensén 2004:533), så er user experience research blevet et almindeligt værktøj for mange ordbogsredaktioner.

Opfordringer til at få brugerens perspektiv i spil er velkendte i leksikografien, og helst med fokus på, som Robert Lew formulerer det: "how actual users use their actual dictionaries in as near natural settings as possible" (Lew 2015:32). Det foreslås også i Atkins & Rundells *The Oxford Guide to Practical Lexicography* fra 2008.

På Jysk Ordbogs (herefter JO) hjemmeside kan man læse at "Jysk Ordbog henvender sig til alle brugere med interesse for de jyske dialekter fra ca. 1700 til ca. 1930 eller for landbefolkningens kultur- og samfundsforhold i samme periode (før landbrugets industrialisering)" (JO's hjemmeside, sektionen Velkommen til ordbogen). Her kan man også se at "fagfolk som sprogforskere, historikere og museumsfolk vil kunne anvende ordbogen som indfaldsvej til relevante emner" (ibid.). Dermed kan man konstatere at JO sigter efter at ramme to ret forskellige grupper: både den alment interesserede og fagpersonen skal have gavn af ordbogens indhold.

JO har ikke altid været helt klar i mælet omkring sin intenderede målgruppe og det polyfunktionelle sigte som den har. Og meget har ændret sig i de 90 år siden professor Peter Skautrup gav sig i kast med det omfattende ordbogsprojekt. Redaktionsreglerne var i 70'ernes prøvehæfter rettet mod et specialiseret publikum i form af filologer, og er siden, først i 80'erne, dernæst løbende blevet tilpasset af flere omgange, fx med enklere lydskrift og forklaringer til atlaskortene, til et mere bredt publikum (jf. Hansen 2020). Med de tanker in mente blev brugerundersøgelsen for JO søsat i vinteren 2020.[1]

Arbejdet tog udgangspunkt i tre overordnede spørgsmål i brugerundersøgelsen af JO:

1) Hvem er JO's målgruppe af brugere anno 2020?
2) Får brugerne gavn af de mange detaljer i ordbogen som redaktørerne lægger så mange kræfter i?
3) Og er jyskordbog.dk udfærdiget på en måde så brugeren får det størst mulige udbytte af besøget på hjemmesiden?

1 Tak til Inger Schoonderbeek Hansen for stort engagement med planlægningen af undersøgelsen, vores samarbejde om oplægget på konferencen samt værdifulde indspark til denne artikel, og til Kristoffer Friis Bøegh for konstruktiv sparring.

I denne artikel vil jeg præsentere brugerundersøgelsen for JO, som blev udført i 2020-2021. I afsnit 2 gennemgås metoderne som bruges i undersøgelsen, JO's forventede målgrupper, samt respondenterne som indgik i undersøgelsen. I afsnit 3 præsenteres resultater og anbefalinger fra projektet. I afsnit 4 diskuteres undersøgelsen og eventuelle fejlkilder, og undersøgelsens resultater afrundes. Artiklen tager afsæt i de data og resultater som er samlet i Svendsen 2021a og 2021b, som er upubliceret materiale, og dermed er dette den første afrapportering af brugerundersøgelsen i leksikografisk regi.

2. Brugerundersøgelse for JO 2021

JO er et historisk dokumentationsprojekt som dækker dialekterne vest for Kattegat, Samsø Bælt og Lillebælt (jf. JO's hjemmeside, sektionen Hvad dækker ordbogen?). I 1932 oprettede professor Peter Skautrup Institut for Jysk Sprog og Kulturforskning hvis hovedopgave var at indsamle materiale til redigering af en arvtager til Feilbergs *Bidrag til en ordbog over jyske almuesmål* (1886-1914). I dag føres det omfattende redaktionsarbejde videre af redaktører på Peter Skautrup Centret for Jysk Dialektforskning (PSC).

JO ligger frit tilgængeligt på jyskordbog.dk og har eksisteret online siden 2000. Ordbogen byder på et væld af informationer og visuelle hjælpemidler hvis formål er at gavne fagfolk og alment interesserede som besøger ordbogen.

Dataindsamlingen til undersøgelsen fandt sted i vinteren 2020-foråret 2021, og databehandling og den efterfølgende rapport blev udarbejdet i foråret 2021. I brugerundersøgelsen var målet ca. 50 besvarelser. Det antal ville være realistisk ud fra både emnet for undersøgelsen samt længden på spørgeskemaet.

2.1. Metode og respondenter

I undersøgelsen har jeg brugt *purposive sampling*-metoden, også kaldet judgement sampling eller formålssampling (jf. Krug & Schlüter 2013:70, Chen 2017:126-127) med udgangspunkt i et spørgeskema. Idéen med formålssampling er at udvælge typer af deltagere som redaktionen ønsker at studere i undersøgelsen, og derudfra opsøge et antal deltagere som

falder inden for de ønskede kategorier (Tagliamonte, som citeret i Krug & Schlüter 2013). Ligeledes benyttedes snowball-metoden (jf. Krug & Schlüter 2013:70) til at få spørgeskemaet ud til et større netværk af mulige respondenter, som også ville være relevante i undersøgelsen. Ved brug af snowball-metoden sendes et spørgeskema til en række respondenter, hvor respondenterne bedes om også at videreformidle spørgeskemaet til deres eget netværk.

Spørgeskemaet kombinerede spørgsmål af kvantitativ og kvalitativ art. At lave en tænke-højt-test eller andre face-to-face-baserede metoder med tilhørende omfattende databehandling (jf. Krug & Schlüter 2013:71-72) var desværre for ressourcekrævende. Derfor faldt valget på at lave et kvalitativt orienteret og dybdegående spørgeskema som gav respondenterne rig mulighed for at komme med input til redaktionen. Et mere omfattende spørgeskema kræver også mere af sit publikum, og derfor var det forventet at ikke alle respondenter ville besvare samtlige spørgsmål i spørgeskemaet. Respondenterne blev også spurgt om typiske baggrundsinformationer såsom alder, beskæftigelse, hjemstavn. Hertil var der også spørgsmål om respondenten brugte JO, og i så fald hvor ofte.

Spørgeskemaet rummede fire opgaver som respondenten kunne springe over, jf. Svendsen 2021a: bilag 1. Den første opgave var multiple choice og havde primært en orienterende funktion. De resterende tre opgaver havde fritekstbesvarelse. Respondenterne blev stillet en opgave og blev bedt om at sætte ord på hvad de gjorde for at løse opgaven. De tre fritekstopgaver er konstrueret på samme måde som en opgave kunne stilles i en tænke-højt-test. I stedet for at respondenten løbende deler sine tanker med intervieweren, blev respondenten i spørgeskemaet bedt om at skrive sine tanker ned til opgaven, både i forhold til hvordan vedkommende løste opgaven, og hvilke problemer vedkommende stødte på undervejs. Da brugbarheden af svarene til opgaverne afhænger af selvrapportering fra respondenten, er der selvfølgelig information som kan være mere eller mindre ubevidst udeladt, og som ville være opdaget og noteret ved en interviewsituation som en tænke-højt-test.

I den første fritekstopgave, *bom*-opgaven, blev respondenten bedt om at finde ud af hvad *bom* betyder i en specificeret sætning. Formålet med opgaven var at undersøge om respondenten kunne navigere mellem homograferne på jyskordbog.dk, i dette tilfælde *bom*1 og *bom*2, begge substantiver. I den anden fritekstopgave, *bøjl*-opgaven, blev responden-

ten bedt om at finde betydningen af *bøjl* i en given kontekst. Dertil skulle respondenten svare på hvor man sagde *bøjl*, og hvad det hedder i flertal. Redaktionen ønskede at undersøge om respondenten kunne finde ud af at folde de relevante afsnit i artikelvisningen ud. Kunne de finde frem til de geografiske og bøjningsrelevante oplysninger uden problemer? I den tredje fritekstopgave, *kaw*-opgaven, skulle respondenterne finde betydningen af ordet *kaw* i Jeppe Aakjærs gendigtning af Robert Burns' venskabsdigt Auld lang syne, Skuld gammel venskab rejn forgo. *Kaw* findes ikke i JO, men *kav* gør da det er den normaliserede form. Formålet var at undersøge: Kan brugerne finde ud af at normalisere deres søgeudtryk når de bruger ordbogen, dvs. *kaw* → *kav*, og får de nok hjælp fra ordbogens "mente du"-funktion til at komme frem til det korrekte svar hvis de ikke normaliserer formen?

Herefter blev respondenterne spurgt ind til deres tanker om hjemmesiden i forhold til udseende, og her var der også et forslag til et nyt layout som respondenterne kunne forholde sig til. De blev også spurgt ind til informationsmængden og de hjælpemidler som jyskordbog.dk indeholder, dvs. især de dialektgeografiske kort. Til de afsluttende holdningsspørgsmål om hjemmesidens udseende, layout, brugervenlighed osv. benyttedes en Likert-skala fra 1-5, se eksempel på figur 1 (i afsnit 3.1. nedenfor).

De kvalitative svar blev behandlet ved brug af manuel kodning. Ved større datamængder vil det være oplagt at bruge computer-assisted qualitative data analysis software (CAQDAS) til databehandling.

2.2. Respondenterne

Ud fra den nævnte formålssampling blev målrettede henvendelser til et udvalg af JO's forventede målgrupper udsendt: leksikografer, lingvister, historikere, museumsfolk, lokalhistorikere og sprogstuderende. Henvendelserne var primært til danske modtagere, men strakte sig også ud over Danmarks grænser til nogle kolleger i Norge og Sverige. Håbet var at redaktionen med hjælp af snowball-metoden kunne få flere besvarelser ind fra de relevante netværk der omgav de enkelte modtagere.

Redaktionen modtog 51 besvarelser i vores dataindsamling med enten delvist eller fuldt besvarede spørgsmål. Fordelingen af respondenternes beskæftigelse kan ses i tabel 1.

TABEL 1. Respondenternes beskæftigelsesmæssige baggrund (Svendsen 2021a:10).

Beskæftigelse (muligt at angive flere pr. person)	**Antal**
Underviser	2
Historiker	5
Lingvist	3
Leksikograf	10
Studerende	13
Pensionist	16
Andet (herunder bl.a. it-udviklere, en slægtsforsker og en museumsinspektør)	7

Gennemsnitsalderen for vores 51 respondenter var 52 år. Den yngste respondent var 22 år, mens den ældste var 85 (se Svendsen 2021a: bilag 3).

3. Anbefalinger og resultater

Brugerundersøgelsen gav brugbare input til en lang række aspekter af JO's indhold, udseende og funktionalitet. Hovedtrækkene kan samles i fire overordnede anbefalinger baseret på resultaterne af undersøgelsen (Svendsen 2021b:7):

1. JO bør få et mere nutidigt design, især med henblik på layout og struktur
2. Søgefunktionen bør optimeres
3. Hovedbetydningerne i ordbogsartiklerne bør være udfoldet som udgangspunkt
4. JO skal være mere synlig: mange ved desværre ikke at den er frit tilgængelig online.

Til de primære anbefalinger følger også en række sekundære anbefalinger, som kort sammenfattes her (Svendsen 2021b:8):

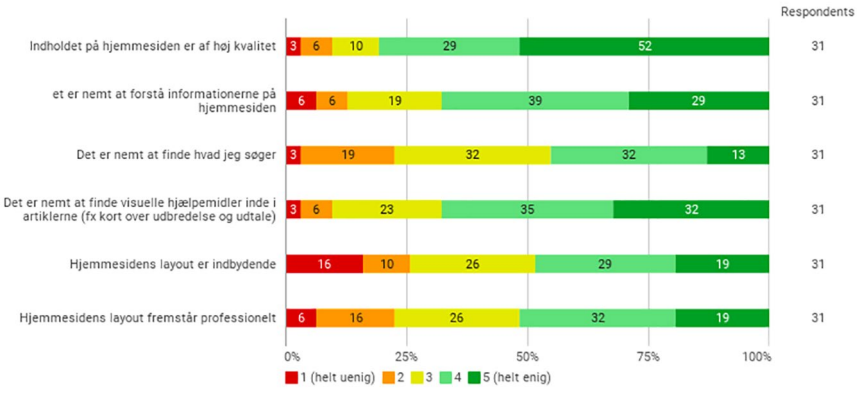
1. En tutorial eller introduktion på hjemmesiden kan hjælpe brugerne med at få overblik over og et større udbytte af JO
2. Flere ord i ordbogen
3. Lettere adgang til redaktionsprincipperne og viden om dialekterne

4. Udtale tilknyttet ordbogsartiklerne
5. Optimal browservisning af JO
6. Spaltevisning med kort bør få en mere glidende overgang – og kortene bør kunne lukkes igen
7. Mere udførlige eller lettilgængelige forklaringer på kortene
8. Interaktive, digitale kort som brugerne kan klikke sig rundt i
9. Flertalsformens nuværende visning bør genovervejes.

I det følgende afsnit uddybes nogle af de tiltag som hører til de primære anbefalinger.

3.1. Layout og struktur

Flere af undersøgelsens respondenter foreslår fornyelse af ordbogens hjemmeside for at JO skal fremstå mere professionel; layoutet er ikke optimalt med dets nuværende udseende. Fx skriver en respondent at JO "virker lidt som om den er lavet i [80'erne]. Meget gammeldags layout og ser ikke lige så professionelt ud, som jeg forventer menneskene bag hjemmesiden er" (Svendsen 2021a: bilag 9). 51 % af respondenterne erklærer sig enige eller helt enige i at hjemmesidens layout fremstår professionelt, og 48 % mener at layoutet er indbydende, se figur 1.



FIGUR 1. Brugernes tanker om det nuværende layout på jyskordbog.dk (Svendsen 2021a:23).

Respondenterne blev i spørgeskemaet præsenteret for et alternativt layout til jyskordbog.dk, et forslag udarbejdet af JO-redaktionen. Her var respondenterne mere positivt stemte, se tabel 2. De havde ligeledes konstruktiv kritik af det nye layoutforslag, fx med henblik på bedre typografi og billedmateriale (Svendsen 2021b:11).

TABEL 2. Respondenter (i %) som har erklæret sig enige eller helt enige i de to udsagn ift. det nuværende layout og layoutforslaget i spørgeskemaet (Svendsen 2021a:25).

Udsagn	**Nuværende layout**	**Nyt layoutforslag**	**Difference (i procentpoint)**
Layoutet er indbydende	48 %	61 %	13
Layoutet fremstår professionelt	51 %	71 %	20

Flere af respondenterne foreslår også at ordbogens forkortelser i højere grad burde opløses. Til en af opgaverne skriver en respondent: "Udbredelsen er tilsyneladende større end blot NVJy, men det er svært at afkode oplysningerne/forkortelserne" (Svendsen 2021b:12). Her kunne JO med fordel kigge til kollegerne på Den Danske Ordbog (DDO) og den løsning de har valgt, nemlig en mouse-over til forkortelserne (som vil virke på mobiltelefoner med et enkelt klik på teksten). Det ville udgøre et stærkt hjælpemiddel for brugerne, og samtidig ikke ændre på længden af og arbejdsmængden til artiklerne.

3.2. Optimering af søgefunktionalitet

At en optimering af søgefunktionen ville være et gavnligt tiltag er synligt i flere dele af respondenternes feedback, men især i den frivillige *kaw*-opgave, som dette afsnit tager udgangspunkt i. Opgaven belyste flere problemstillinger som redaktionen var interesseret i, se afsnit 2.1. Opgaven viste tydelige problemer for respondenterne, se tabel 3.

TABEL 3 (Svendsen 2021a:21). Oversigt over kodning af svar fra *kaw*-opgaven.

Resultat fra kodning af respondenternes opgaveløsning	**Antal respondenter**
Finder det rigtige svar "venstre (hånd)", enten som sb. eller adj. i ordbogen	7
Finder svar andet sted end ordbogen (fx via Google) eller gætter på det	6
Finder ikke frem til det korrekte svar	15
Svar der ikke kan afkodes	1
I alt	**29**

Fritekstsøgning vil være et gavnligt tiltag på jyskordbog.dk, som på mange måder ville være oplagt for en dialektordbog. Flere respondenter påpeger dette i løsningen af *kaw*-opgaven, fx skriver en respondent: "Eg prøvde først med fritekstsøk, men det ser ikkje systemet ut til å støtte?" (Svendsen 2021a:22). Med fritekstsøgning ville brugerne kunne finde citater med den pågældende stavemåde med det samme, uden at skulle være bevidste om at normalisere formen først. I forhold til normalisering er den nuværende søgefunktion heller ikke til stor hjælp: Ved søgning på formen *kaw* får man hele 20 forslag i "mente du"-funktionen, og kun det 13. forslag i rækken fører respondenten videre til det korrekte opslag (jf. Svendsen 2021b:16–17).

3.3. Udfoldede betydningsafsnit

Flere respondenter fremhæver både i den indledende orienteringsopgave og i *bøjl*-opgaven at det ville være en hjælp hvis betydningerne i artikelopslagene var udfoldede som udgangspunkt, fx skriver en respondent: "Et problem er at man skal folde betydninger ind og ud (eller først indstille artiklen til at være fuldt udfoldet), hvilket langsommeliggør søgningen" (Svendsen 2021b:14). Hvis man kigger i store onlineordbøger som Svenska Akademiens ordbok (SAOB) og DDO ses det også at den gængse form er en udfoldet betydning: Det tyder på at brugerne foretrækker at skulle scrolle lidt for at finde svar, da det er et format de kender i forvejen.

3.4. Synlighed

En respondent skriver at "Jeg kendte ikke Jysk Ordbog i forvejen, men jeg vil bestemt benytte mig af den i fremtiden" (Svendsen 2021b:18). Denne kommentar viser tydeligt et af JO's store problemer: Hjemmesiden ligger godt gemt selvom den er frit tilgængelig online (jf. Svendsen 2021b: 18). En af de vigtigste årsager er den tekniske opsætning af hjemmesiden som gør at hele ordbogen og samtlige opslag har den samme URL. Derfor kan man ikke linke til specifikke opslagsord, hvilket ellers er en af de mest almindelige funktioner i en online-ordbog. Dette betyder også at jyskordbog.dk 'drukner' i en Google-søgning, for webcrawlerne får meget lidt at arbejde med. En søgemaskineoptimering, som hjemmesiden også vil have gavn af, er først relevant efter at dette grundlæggende problem med opsætningen er løst.

En af respondenterne forslår at "i ei ideell verd skulle det ein stad vore tilvising til tilsvarande artiklar i dei andre store ordboksverka i Norden, såleis til Ømålsordbogen, ODS, til *dugurd* (Norsk Ordbok, Norsk Riksmålsordbok), *dagvard* (SAOB)". Dette stemmer godt overens med JO-redaktionens egne tanker om mulighederne for en samlet dansk dialektportal, og det er positivt at der er interesse i forhold til at kunne linke flere nordiske ordbøger sammen.

4. Fejlkilder og diskussion

Dataindsamlingen til brugerundersøgelsen fandt sted i en periode hvor mange af Danmarks kulturinstitutioner og arbejdspladser var ramt af nedlukning pga. COVID-19. Dette kan have haft betydning for det relativt lille antal besvarelser fra især museer samt lokalhistoriske foreninger og arkiver: Museerne var lukkede, og hvis mange af medarbejderne fra de lokalhistoriske foreninger og arkiver har været frivillige, kan det ikke forventes at de arbejdede hjemmefra under nedlukningen (Svendsen 2021a:33).

Flere respondenter gik i stå ved de frivillige opgaver (se Svendsen 2021a: bilag 1). Selvom det var et forventeligt minus ved at lave et spørgeskema der krævede mere af sine respondenter er udfordringen værd at være opmærksom på ved en senere undersøgelse (Svendsen 2021a:34). Spørgeskemaets form tog udgangspunkt i at respondenter med en del såvel som ingen erfaring med ordbogen skulle kunne besvare spørgsmålene. Der-

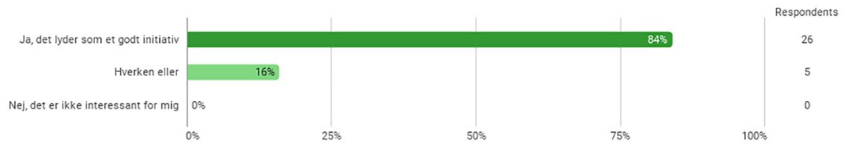
for kom opgaveløsningen før spørgsmålene til layout, indhold m.v., da de førstegangsbesøgende dermed fik noget viden om sidens opbygning som de kunne tage udgangspunkt i. Færre opgaver ville med stor sandsynlighed resultere i flere gennemførte besvarelser fra respondenterne, men det er en afvejning af kvantitet og kvalitet, som det ofte er tilfældet (Svendsen 2021a:34).

4.1 Perspektiver og afrunding

Lad os vende tilbage til udgangspunktet for undersøgelsen, nemlig vores indledende overvejelser om målgrupper og brugervenlighed præsenteret i afsnit 1. Dataene fra projektet viser at respondenter som havde angivet deres beskæftigelse som leksikograf eller studerende, gav konstruktivt input ved fritekst-besvarelser. Andre grupper som fx historikerne og de der angav et museumsrelevant fag, kom med meget begrænset input til fritekst-besvarelser i undersøgelsen. Kræves der for specifikt leksikografiske eller lingvistiske kompetencer til at fx historikere og museumsfolk (jf. afsnit 1) får det fulde udbytte af ordbogen (Svendsen 2021b:26)?

Den alment interesserede bruger møder ifølge brugerundersøgelsens resultater en ordbog for især lingvistisk orienterede fagfolk, som bør være mere imødekommende og tilgængelig for et større publikum, hvis ordbogen fortsat ønsker at ramme den gruppe. Værdifulde tiltag vil være justeringer af layout og struktur så det svarer til hvad de forventer af en onlineordbog, samt lydlige elementer der bevirker at dialekterne ikke kun læses, men også kan høres i et vist omfang, se figur 2 (Svendsen 2021b:26).

I forhold til detaljegraden forsøgte redaktionen også at få nogle svar på hvilke områder redaktørerne fremadrettet ville kunne skære ned på, fx i forhold til etymologi og udbredelseskort. Her stødte redaktionen imidlertid på en ofte hørt respons: Ordbogsbrugere vil, når de bliver spurgt, ofte gerne have mere af det hele og vil sjældent vurdere at noget er overflødigt. Respondenterne udtrykte begejstring for det store tilgængelige materiale, og de mange hjælpemidler som kort og kilder får positiv respons. Dog bør hjælpemidlerne tilpasses for at redskaberne bliver nemmere at bruge (jf. Svendsen 2021b:25, Svendsen 2021a:27).



FIGUR 2. Respondenternes holdning til lydlige tiltag fra JO (Svendsen 2021a:26).

Brugerundersøgelsen giver et godt grundlag for at JO fremadrettet kan evaluere funktion, udseende og brugervenlighed. Efter undersøgelsen har PSC søsat en podcast med korte, informative afsnit om jysk sprog, kultur og historie. Centrets hjemmeside, jysk.au.dk, er blevet moderniseret og har fået et interaktivt kort med dialektlydprøver. PSC arbejder på at søge infrastrukturmidler til at implementere anbefalingerne fra brugerundersøgelsen på jyskordbog.dk, og yderligere til at digitalisere JO's seddelsamling, som også kræver ændringer i hjemmesidens struktur og søgefunktion, jf. Hansen, Svendsen & Bøegh (2023).

Lad os slutte på to opløftende kommentarer fra respondenterne: En tidligere leksikograf skriver at "det er en velstrukturert og faglig grundig dokumentasjon av jysk ordforråd og ordbetydninger. Den utgjør et viktig grunnlag for skandinaviske komparative dialekt- og ordstudier," og en studerende bidrager med at det er "sjovt at man kan slå jyske ord op! Et godt værktøj, der ikke er svært at håndtere" (Svendsen 2021b:27).

Litteratur

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

Chen, Qiongqiong 2017. Appendix A: Methodology and Methods. *Globalization and Transnational Academic Mobility: The Experience of Chinese Academic Returnees*. Singapore: SS+B Media Singapore and Higher Education Press.

DDO = *Den Danske Ordbog, udgivet af Det Danske Sprog- og Litteraturselskab*. 1991-. <ordnet.dk/ddo>. Besøgt august 2022.

Feilberg, Henning Frederik 1886-1914. *Bidrag til en ordbog over jyske almuesmål*. København: Universitets-jubilæets danske samfund.

Hansen, Inger Schoonderbeek 2020. Jysk Ordbog – af hvem, til hvem og hvorfor? I: Sandström, Caroline, Ulla-Maija Forsberg, Charlotte af Hällström-Reijonen, Maria Lehtonen og Klaas Ruppel (red.), *Nordi-*

ska studier i lexikografi 15. Helsingfors: Nordiska föreningen för lexikografi, 135–143.

Hansen, Inger Schoonderbeek, Mette-Marie Møller Svendsen & Kristoffer Friis Bøegh 2023. Digitalisering af Jysk Ordbogs seddelsamling. I: Holmer, Louise et al. (red.), *Nordiska studier i lexikografi 16*. Lund & Göteborg: Nordiska föreningen för lexikografi, 99–112.

JO = *Jysk Ordbog, udgivet af Peter Skautrup Centret for Jysk Dialektforskning*. 1970-. <jyskordbog.dk>. Besøgt august 2022. Herunder sektionerne: Velkommen til ordbogen, Hvad dækker ordbogen?

Krug, Manfred & Julia Schlüter (red.) 2013. *Research Methods in Language Variation and Change*. New York: Cambridge University Press.

Lew, Robert 2015. Dictionaries and their users. I: Hanks, Patrick & Gilles-Maurice de Schryver (red.), *International Handbook of Modern Lexis and Lexicography*. Berlin, Heidelberg: Springer, 1–9.

ODS = *Ordbog over det danske Sprog, udgivet af Det Danske Sprog- og Litteraturselskab*. 1918-1956. <ordnet.dk/ods>. Besøgt august 2022.

SAOB = *Svenska Akademiens ordbok*. 1898–. <saob.se>. Besøgt august 2022.

Svendsen, Mette-Marie Møller 2021a. Brugerundersøgelse for Jysk Ordbog: En kvantitativ og kvalitativ undersøgelse af brugernes oplevelse og tilfredshed med Jysk Ordbogs hjemmeside. Rapport og datasæt. Ikke-publiceret materiale.

Svendsen, Mette-Marie Møller 2021b. Brugerundersøgelse for Jysk Ordbog: En kvantitativ og kvalitativ undersøgelse af brugernes oplevelse og tilfredshed med Jysk Ordbogs hjemmeside. Præsentationshæfte. Ikke-publiceret materiale.

Svensén, Bo 2004. *Handbok i lexikografi. Ordböcker och ordboksarbete i teori og praktik*. 2 uppl. Stockholm: Norstedts Akademiska Förlag.

Automatic Terminology Extraction: New Challenges in Terminology Work in Iceland

Ágústa Þorbergsdóttir, Atli Jasonarson, Finnur Ágúst Ingimundarson, Einar Freyr Sigurðsson, Steinþór Steingrímsson & Hjalti Daníelsson

We present TermPortal, a web-based terminology acquisition and management system to support Icelandic terminology work. We discuss its function and report on how it copes with a selected domain-specific field, namely linguistic terms. Two different automatic extraction methods of building a terminology are explored: The first utilizes a tf-idf (term frequency-inverse document frequency) model and returns tokens with a statistics-based score; if it is higher than a certain threshold, a probable technical term is suggested. The second method makes use of Daðason's BiLSTM Compound Splitter for Icelandic (Kvistur), focusing on the different morphological parts which a given term can consist of. We also try combining these two methods.

KEYWORDS: terms, terminology extraction, domains, Icelandic

1. Introduction

Manual collection of terminology is a very time-consuming task.[1] Terminology work in Iceland has a long history and the work is usually carried out by terminology committees. Over 50 terminology committees of various kinds have existed in Iceland for longer or shorter periods, as regards activity and working methods. Usually, those committees are composed of subject-matter experts, working on a voluntary basis, who have devoted their time to manually creating glossaries within their field. Improving the process of building a new terminology — or finding new words to add to an existing one — and speeding it up is therefore of

1 The work presented in this paper was supported by Rannís (Strategic Research and Development Programme; grant #180017-53011). We thank two anonymous reviewers and the editors for their comments. We also thank Eiríkur Rögnvaldsson whose chapter in his open-access book, *Hljóðkerfi og orðhlutakerfi íslensku*, provided the basis for our case study discussed in sections 3 and 4.

utmost importance and very useful for standardizing vocabulary in specialized fields.

This paper introduces the web-based terminology acquisition and management system TermPortal,[2] which supports Icelandic terminology work. It is intended to facilitate lexicographic work in building terminologies. TermPortal helps us identify terms and see how they are used, which is important, e.g., when defining concepts and the boundaries of LSP (language for special purposes) vs. LGP (language for general purposes).

The paper is structured as follows: Section 2 provides an overview of TermPortal and how texts are processed by automatic term extraction tools. Section 3 presents a case study on linguistic terms to test TermPortal's function and precision while section 4 reports on TermPortal's evaluation. Section 5 presents conclusions of the study and future work.

2. TermPortal

TermPortal is a terminology acquisition and management system. TermPortal consists of two main parts. Firstly, the TermPortal workbench includes an automatic pipeline to extract terminology from media and a web platform where users can create, manage and maintain termbases. Secondly, the automatic term extraction (ATE) system is a central component in the TermPortal workbench but can also be used independently. We looked beyond the traditional methods of manual terminology work and tried to simplify the process of preparing, storing, and sharing terminology glossaries.

Users of the system can upload texts which are then processed by ATE tools, tagging potential terminology candidates for the user to accept or decline. The process is as follows: The user uploads a text file and — optionally — specifies a domain, such as medicine, history or linguistics. TermPortal can use termbases for different fields or run without any support from a termbase. The text is then run through a pipeline, consisting of six steps:

1) The text is tokenized into single-word units, and its punctuation marks are removed.

2 https://termportal.arnastofnun.is/

2) The tokenized text is run through a part-of-speech tagger (with ABL-Tagger as default; see Steingrímsson et al. 2019), returning every token along with its corresponding tag.
3) The tags are used to remove unwanted words, such as foreign ones, proper nouns and numbers, as well as single-character units.
4) The tokens are lemmatized using Nefnir (Ingólfsdóttir et al. 2019), whose accuracy improves substantially when supplied with part-of-speech tags.
5) The lemmatized tokens are run through a tf-idf (term frequency-inverse document frequency) model, trained on roughly 17 million words collected from scholarly and scientific journals and websites (Barkarson et al. 2021).
6) At this stage, there are three options:
    a) The tokens whose tf-idf score is above a given threshold are returned as probable technical terms.
    b) The tokens can be split into their stem structure. If a list of known stems and morphological parts exists for a given field, it can be used to identify tokens as probable terms. If, for example, a stem which is a part of a compound is on such a list, the whole compound may be suggested as a term.
    c) These two methods can be used individually or combined.

The tokens the system identifies as probable technical terms are returned to the user via the web interface. The user is then faced with their text where candidate terms are highlighted, see Figure 1. The user is asked to accept or reject the candidates and their decisions are subsequently stored. Previously accepted terms or terms that are in an available termbase are highlighted in green, while unknown candidate terms are highlighted in blue. The accepted terms are added to the termbase and the rejected ones, highlighted in red, are added to a list of candidates rejected as terms in this domain. The rejected ones are not, however, entirely discarded as they can be used to filter out irrelevant suggestions in future use. The accepted terms are also useful, as different morphological parts can serve to expand the collection and therefore improve the method described in 6b). Additionally, the user can, anywhere in the process,

add their own words, such as ones the system missed or the ones it mis-lemmatized.

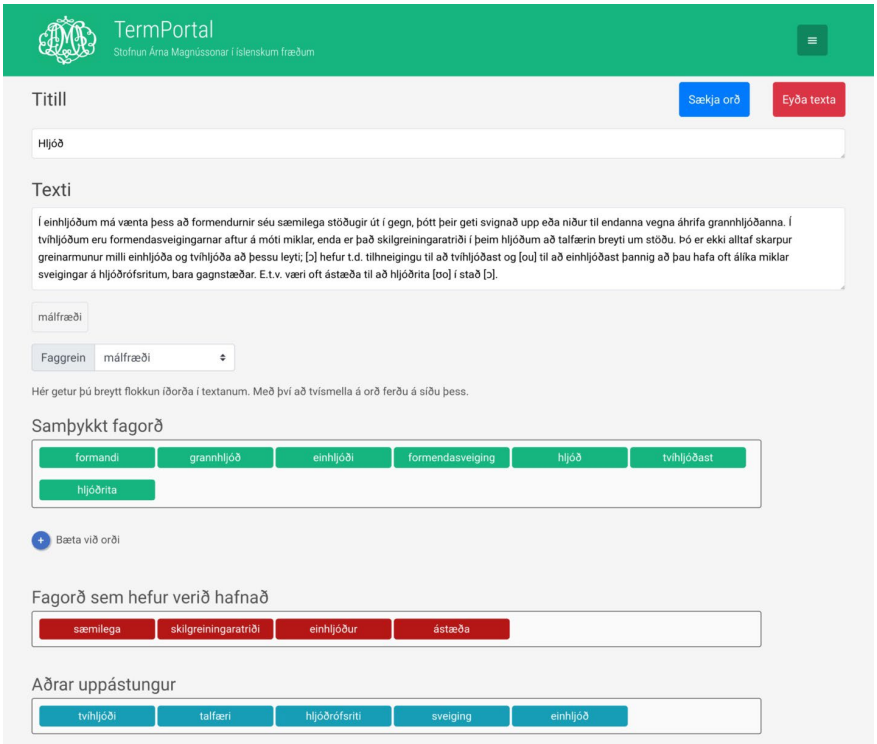


FIGURE 1. A screenshot of TermPortal's web interface in which the user decides which suggestions to accept or reject.

This pipeline — see Figure 2 — only identifies single-word units. We have also experimented with multi-word unit term extraction applying three different methods for identifying term candidates: C-value (Frantzi et al. 2000); an approach based on stem ratio (Daníelsson et al. 2020); and Levenshtein-distance (Levenshtein 1966) between possible candidates and known terms. These approaches are described in more detail in Daníelsson et al. (2020). While these approaches are effective in finding the terms, they also produce many false positives, leading to the effect of the results not being very helpful in boosting productivity of termbase editors. Efforts to raise the accuracy of multi-word term extraction are not within the current scope of our project although we want to study that in future work.

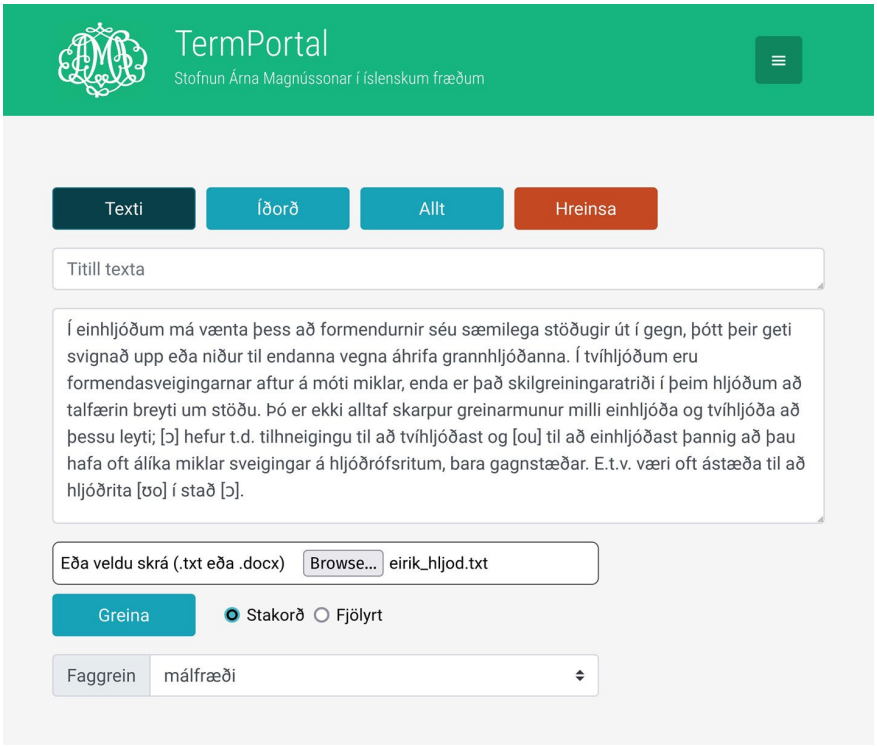


FIGURE 2. A screenshot of the entry point of TermPortal's web interface.

Below the text area in Figure 2 is the analysis button (Icel. *greina*) and two options allowing the user to either extract single-word terms (Icel. *stakorð*) or multi-word terms (Icel. *fjölyrt*), with the latter option still being at an experimental stage. At the bottom, the domain (Icel. *faggrein*) is specified, in this case grammar (Icel. *málfræði*).

3. A case study on linguistic terms

To test and evaluate TermPortal's function and precision, applying the methods discussed in the previous section, we used a chapter on Icelandic speech sounds (chapter 4) from Rögnvaldsson's (2013) book on Icelandic phonology and morphology. The chapter is 16 pages long and it took a single person in the project two hours to mark linguistic terms in the text. The manual registration of terms in the text resulted in 99 terms, which we use as a gold standard (see discussion on evaluation in section 4 below).

One particularity of texts on phonology such as the one chosen for this paper is that some of the terms can be lined up together to form another multi-word term, which raises the question whether that new term should be considered as a sum of its parts and therefore a unique term. A case in point could for example be the classification of vowels in Icelandic, which are divided into either front or back, high or low, or rounded or unrounded vowels. They can furthermore either be long or short. A combination of these terms could therefore be used to describe a certain vowel, for example the sound [i] which is unrounded and the most front and highest vowel in Icelandic, i.e., *frammælt*, *nálægt*, *ókringt sérhljóð* 'front, high, unrounded vowel'. In the present data each adjective was defined and marked individually and combinations such as the example shown here are not considered as terms on their own.

Another classic problem of defining the boundaries of LSP (language for special purposes) vs. LGP (language for general purposes) or simply terminology vs. general vocabulary concerns the use of ordinary adjectives, nouns, etc. as specific terms, as for example the adjective *nálægur*, which has the general meaning 'near(by), close', whereas in the preceding paragraph it has a specific function as a term to describe a certain vowel quality (Eng. *high*). One could also point to the distinction of terms relating to the speech organs, such as between the very much ordinary nouns *tongue* and *lips* as opposed to *palate*, *dorsum linguae* and *uvula*.

4. Evaluation

To evaluate TermPortal's suggestions, the list of 99 terms acts as a 'ground truth' or gold standard, meaning that a perfect model would extract all the 99 terms, and nothing else. That would yield an F_1 score of 1 (or 100%) which is computed by calculating the following:

- True positives (TP): All the terms extracted from the text that are present in the ground truth
- False positives (FP): All the terms extracted from the text that are not present in the ground truth
- False negatives (FN): All the terms not extracted from the text that are present in the ground truth

When these numbers have been calculated, the recall (R) is computed by the equation $\frac{TP}{TP+FN}$, i.e., true positives divided by all terms present in the ground truth. On its own, recall is not very useful, though, because a model could achieve recall of 100% by simply returning all the possible elements. Therefore, precision (P) is used to calculate how accurate the output of the model is: $\frac{TP}{TP+FP}$. The F_1 score is the harmonic mean of these two, which gives us a single number to represent how effective a classifier, such as the one in question, is. It is computed as follows: $2 \times \frac{P \times R}{P+R}$.

As mentioned in section 2, two different automatic methods were taken into consideration and compared to see which one proved to be more effective as a classifier for TermPortal. The first one was based on a tf-idf (term frequency-inverse document frequency) method, running the lemmatized tokens from the text through a tf-idf model specially trained on scholarly and scientific material, returning tokens with a score higher than a certain threshold as probable technical terms, i.e., essentially words considered 'rare' in the model. The output was a diverse list which included amongst other things all or most of the phonetically transcribed words appearing in the text, such as [kʰaltʏr] for *kaldur* 'cold', i.e., not words to be considered as terms. As a result, the efficiency, i.e., the F_1 score, was accordingly rather low and the numbers read as shown in Table 1.

TABLE 1. A method based on tf-idf

True positives:	46
False positives:	65
False negatives:	53
Precision:	41.4%
Recall:	46.5%
F_1:	0.438

The other approach was based on a BiLSTM Compound Splitter for Icelandic (Kvistur) (Daðason et al. 2020) which was used on linguistic terminology from *The Icelandic Term Bank* (https://idord.arnastofnun.is/),

1,367 terms in all. Compounds are extremely common in Icelandic[3] and that also goes for terms where different morphological parts, stems in particular, serve as essential building blocks. As an example, we can take the Icelandic name of the term bank, *Íðorðabankinn*, which we would expect Kvistur to split into the following parts: *íð* 'work, profession', *orða* 'words' and *bankinn* '(the) bank'.

The resulting list, i.e., from using the compound splitter on linguistic terminology, contained several incorrectly split parts, but these were relatively few. One such example were the parts *for* 'pre-/pro-, mud',[4] and *morð* 'murder' which form a non-existing word (*for.morð*), instead of the correct parts *form* 'form' and *orð* 'word' (*form.orð*), which were, nota bene, also included in the list. All such irregularities were removed from the list, leaving 739 morphological parts, which were used as the basis for identifying terms from the text, retrieving words containing at least one morphological part (for example *form* or *orð* or even both). This gave the results shown in Table 2.

TABLE 2. A method based on a BiLSTM Compound Splitter for Icelandic

True positives:	93
False positives:	150
False negatives:	6
Precision:	38.3%
Recall:	93.9%
F_1:	0.544

As the numbers show, this resulted in a higher F_1 score, with a much higher recall but less precision, i.e., the recall is less accurate.

Comparing the two methods, we can see that the method that uses the compound splitter results in a much higher recall, meaning that a database containing known stems and morphological parts of a given field's vocabulary can be of great help when extracting new ones. So far, this has only been investigated with a single book chapter as test data, but

3 The majority of words in the Database of Icelandic Morphology consist of compounds: "Out of 278,764 paradigms [...] on Dec. 15th 2015, 32,118 entries were non-compounds, and the remaining 246,646 entries were compounds" (Bjarnadóttir 2017:14).

4 In Icelandic *for* can be used as a prefix 'pro, pre, etc.' or a noun meaning 'mud'.

the results are promising. It should be noted, however, that its precision is quite low, 38.3%, meaning it returns multiple false positives, which is disadvantageous as it can be time-consuming for editors to filter out the false positives.

The tf-idf method has neither high recall nor high precision, which stems from the fact that it returns only 111 candidates, compared to the 243 candidates the compound method suggests, and it suggests multiple words that cannot be considered true positives.

Finally, we combined the two methods, which resulted in a list containing only terms deemed probable by both, yielding the results shown in Table 3.

TABLE 3. The two methods referred to above combined

True positives:	46
False positives:	15
False negatives:	53
Precision:	75.4%
Recall:	46.5%
F_1:	0.575

The combined method yields interesting results: The recall is the same as for the tf-idf method, at 46.5%, but its precision, 75.4%, is the highest one, meaning that for every four words the method returns, three of them are correct.

5. Conclusions and future work

We need to look beyond the traditional (and perhaps somewhat dated methods) of manual termbase construction and try to simplify the process of preparing, storing, and sharing term glossaries. The automatic term extraction tool, built for the workbench, shows promising results. As noted, it is the first tool of its kind to support Icelandic, and terminology databases have until now been constructed by hand. As a result, our focus was on maximizing the tool's ability to gather potential new terminology and create a sizable initial database suitable for further computerized work and research. Accordingly, term recall was of primary importance and was heavily emphasized over precision during the tool's

development. Fine-tuning precision will be part of future work on TermPortal.

Furthermore, the numerous available terminology databases for Icelandic should be looked into in more detail and used as a basis for further development of the compound-splitter method. Moreover, once fully functional and voluminous enough, TermPortal's data can be used in various, useful ways, such as automatic indexing of scholarly and scientific work or automatic keyword extraction for all sorts of texts.

References

Barkarson, Starkaður, Steinþór Steingrímsson, Hildur Hafsteinsdóttir, Þórdís Dröfn Andrésdóttir, Inga Guðrún Eiríksdóttir & Bolli Magnússon 2021. *IGC-Journals-21.12* (*The Icelandic Gigaword Corpus – scholarly and scientific journals*), CLARIN-IS. <http://hdl.handle.net/20.500.12537/166>. Accessed August 2022.

Bjarnadóttir, Kristín 2017. Phrasal compounds in Modern Icelandic with reference to Icelandic word formation in general. In: Carola Trips & Jaklin Kornfilt (eds.). *Further investigations into the nature of phrasal compounding*. Berlin: Language Science Press, 13–48.

Daðason, Jón, David Erik Mollberg, Hrafn Loftsson & Kristín Bjarnadóttir 2020. Kvistur 2.0: a BiLSTM Compound Splitter for Icelandic. In: Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.). *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille: European Language Resources Association, 3991–3995.

Daníelsson, Hjalti, Ágústa Þorbergsdóttir, Steinþór Steingrímsson & Gunnar Thor Örnólfsson 2020. TermPortal: A Workbench for Automatic Term Extraction from Icelandic Texts. In: Béatrice Daille, Kyo Kageura & Ayla Rigouts Terryn (eds.). *Proceedings of the 6th International Workshop on Computational Terminology* (COMPUTERM 2020). Marseille: European Language Resources Association, 8–16.

Frantzi, Katerina T., Sophia Ananiadou & Hideki Mima 2000. Automatic recognition of multi-word terms: the *C-value/NC-value* method. *International Journal on Digital Libraries* 3, 115–130.

Ingólfsdóttir, Svanhvít Lilja, Hrafn Loftsson, Jón Friðrik Daðason & Kristín Bjarnadóttir 2019. Nefnir: A high accuracy lemmatizer for Icelandic. In: Mareike Hartmann & Barbara Plank (eds.). *Proceedings of the 22nd Nordic Conference on Computational Linguistics*. Turku: Linköping University Electronic Press, 310–315.

Levenshtein, Vladimir Iosifovich 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady* 10(8), 707–710.

Rögnvaldsson, Eiríkur 2013. *Hljóðkerfi og orðhlutakerfi íslensku*. Reykjavík. <https://notendur.hi.is/eirikur/hoi.pdf>. Accessed September 2022.

Steingrímsson, Steinþór, Örvar Kárason & Hrafn Loftsson 2019. Augmenting a BiLSTM tagger with a morphological lexicon and a lexical category identification step. In: Ruslan Mitkov & Galia Angelova (eds.). *Proceedings of Recent Advances in Natural Language Processing,* RANLP 2019. Varna: INCOMA Ltd., 1161–1168.

Det Centrale Ordregister og dets leksikografiske anvendelser

Thomas Widmann

In this article, we introduce the new Danish language resource called *Det Centrale Ordregister* 'the Central Word Registry' (abbreviated *COR*). It assigns unique and permanent ID numbers to lemmas and word forms in Danish. It consists of three levels; the first one corresponds to the official Danish orthographical dictionary (*Retskrivningsordbogen*), published by the Danish Language Council. The second level will contain resources produced by various professional linguistic organisations, and the third one will be open to everybody who is interested.

We start out by looking at the structure of the COR in some detail. Following on from this, we examine how various historical orthographical dictionaries can be encoded in the COR, and how this can be used to improve the implementation of the Danish Language Council's historical dictionary comparison site, RO^{hist}.

Finally we explore other lexicographic uses of the COR, and we examine how this could be expanded to create links to closely related languages, in particular Norwegian.

NØGLEORD: sprogteknologi, leksikografisk database, retskrivning, sproghistorie

1. Introduktion

Der findes mange elektroniske resurser for dansk – ordbøger i maskinlæsbare formater, korpusser, taggere, etc. – men de kan være svære at genbruge da de ikke er baseret på de samme grundresurser og ikke deler databasenøgler og tilsvarende, og det gør det sværere end nødvendigt at lave sprogteknologi på dansk. Vi har derfor lavet en ny resurse som forsøger at løse dette problem: Det Centrale Ordregister.

Det Centrale Ordregister (COR) tildeler unikke id-numre til alle lemmaer og ordformer på dansk. Det er et projekt under Digitaliseringsstyrelsen, med deltagelse af Dansk Sprognævn, Det Danske Sprog- og Litteraturselskab og Center for Sprogteknologi ved Københavns Universitet. I Dansk Sprognævn er vi ansvarlige for grundregisteret: ortografi og morfologi for det ordforråd som dækkes af Retskrivningsordbogen. Dette grundregister blev lanceret i september 2022 og er tilgængeligt på ordregister.dk

Vi vil i denne artikel først præsentere COR's opbygning, beskrive grundresursens struktur og demonstrere hvordan nye COR-resurser kan tilføjes. Derefter vil vi se på forskellige leksikografiske anvendelser med særligt fokus på RO^{hist}, Dansk Sprognævns hjemmeside som tillader sammenligning af forskellige historiske retskrivningsordbøger. Vi vil dernæst diskutere COR-linkere (programmer der automatisk tildeler COR-id-numre til alle ord i en løbende tekst), og til sidst vil vi som perspektivering se på hvordan man kan forestille sig at det danske COR kan sammenknyttes med et hypotetisk norsk parallelprojekt. Vi håber herved at give læseren både lyst til at begynde at bruge COR og de praktiske færdigheder til at gøre det.

2. Motivation

I Danmark har vi Det Centrale Personregister (CPR) som tildeler CPR-numre eller personnumre (svarende til *fødselsnummer* i Norge, *personnummer* i Sverige, *henkilötunnukset/personbeteckningar* i Finland og *kennitölur* i Island). Det centrale element er disse unikke numre som alle indbyggere har. Det er smart fordi det er en nøgle som tillader forskellige databaser at tale sammen.

Uden en sådan nøgle er databasesammenkøring en besværlig og tidskrævende proces da man typisk skal matche personerne på navn, fødselsdato og adresse; ingen af dem er unikke, og både navn og adresse kan ændre sig over tid.

På det sproglige område har en tilsvarende offentlig databasenøgle manglet. Mange leksikografiske og datalingvistiske projekter har måske tilknyttet unikke id-numre til deres lemmaer, men hvis disse ikke deles af andre projekter, vil det fortsat være besværligt og tidskrævende at køre forskellige databaser sammen så de kan bruges i andre projekter.

Hvis man fx har to ordbøger, en med udtaleangivelser og en anden med betydningsangivelser, kan man ikke umiddelbart flette dem sammen automatisk. Den største del af et sådant projekt er ikke vanskeligt – de fleste lemmaer har kun én udtale og ét bøjningsparadigme – men en automatisk sammenfletning kommer typisk til kort når den møder homografer og homonymer; på dansk fx ord som *kost* (/kɔsd/ "fejeredskab" eller /kʌsd/ "føde"), *tag* ("øverste del af et hus" eller det engelske låneord) eller *frø* (*en* frø "padde" eller *et* frø "plantedel"). Det gør det besværligt og tidskrævende at genbruge sproglige resurser.

Vi prøver derfor med COR-projektet at støtte sprogteknologi på dansk ved at lancere de manglende offentlige databasenøgler i håbet om at mange projekter vil begynde at bruge dem.

Der er paralleller mellem COR og den norske *Metaordboka* (se fx Grønvik & Ore 2018 og Ore et al. 2023); den primære forskel er nok at hovedmotivationen for COR er at dele data og lave resurser som kan arbejde sammen, hvorimod *Metaordboka* i udgangspunktet er et internt værktøj.

3. Hvordan fungerer det?

COR's opdeling af grundordforrådet i lemmaer er baseret på *Retskrivningsordbogen* fra Dansk Sprognævn. Derfor følger det samme princip for hvad et lemma er (RO2012:13f):

> Opdelingen i opslagsord er principielt uafhængig af ordenes betydning. Det bevirker at ord med forskellig betydning er slået sammen i ét opslagsord hvis de i øvrigt har samme stavemåde, udtale, ordklasse og bøjning, og hvis de indgår i sammensætninger på samme måde.

Et COR-nummer vil derfor svare til et lemma (med ordklasse), og alle almindelige bøjningsformer anføres derunder.

Vi kender ikke til noget projekt, hverken i Danmark, Norden eller resten af verden, som er fuldstændigt sammenligneligt med COR. Der findes naturligvis talrige leksikalske databaser med morfologisk information, men det særlige ved COR er den åbne og fleksible struktur med en fælles nøgle som muliggør at man tilføjer alskens resurser (herunder historiske ortografiske ordbøger), sammen med de såkaldte *relationer* (se nedenfor) som tillader at disse mangfoldige resurser peger på hinanden.

COR's id-numre er principielt arbitrære. Grundresursens id-numre ligger mellem 0 og 99.999, og de er ikke tildelt alfabetisk. Af praktiske årsager har vi opdelt dette interval efter ordklasse, fx ligger adjektiverne mellem 15.000 og 29.999, og substantiverne mellem 40.000 og 99.999, men dette er ikke et formelt krav, og andre COR-resurser forventes ikke at følge dette mønster. Som et eksempel på id-numrenes arbitraritet er her de første 15 adjektiver: COR.15006 *dansk*, COR.15021 *travl*, COR.15026 *lille*, COR.15027 *vidunderlig*, COR.15049 *smart*, COR.15052 *alvorlig*, COR.15053 *lækker*, COR.15064 *ny*, COR.15066 *flest*, COR.15067

yderlig, COR.15073 *politisk*, COR.15075 *gylden*, COR.15081 *dyr*, COR.15082 *nær*, COR.15083 *kongelig*.

4. COR's struktur: eksempler

Lad os nu se på et konkret eksempel: *bark*. Det kan på dansk betyde både "det yderste lag på et træ" og "et skib", men det kan kun staves på denne måde, ordklassen kan kun være et substantiv, kønnet er fælleskøn, og udtalen kan kun være /ba:g/. Det er derfor tæt på kun at kræve ét COR-id (da opdelingen i opslagsord som sagt er uafhængig af ordenes betydning). Men der er to bøjningsmønstre: ét med pluralis og ét uden. Det betragtes derfor som to lemmaer, og de tildeles to COR-id-numre, COR.69850 til lemmaet uden pluralis, og COR.36198 til det med.

Da grundformen staves ens ("bark"), tildeler vi også en disambiguerende glosse til de to indgange. Vi har således:

- COR.58005 *sb* (*yderste lag på et træ*)
- COR.59594 *sb* (*sejlskib*)

Til bøjningsformerne bruges der to udvidelser: grammatisk kode og ortografisk variation. Til et substantiv af fælleskøn i singularis ubestemt er den grammatiske kode 110, i singularis bestemt er den 111, og så videre. Ordklassen, som altid er det første element i den grammatiske kode, er helt baseret på *Retskrivningsbogen*. I tabel 1 nedenfor er den ortografiske variation altid 01:

TABEL 1. lemmaet 'bark' i COR.

COR-id	lemma	Glosse	gram. funk.	form
COR.58005.110.01	bark	yderste lag på et træ	sb.fk.sg.ubest	bark
COR.58005.111.01	bark	yderste lag på et træ	sb.fk.sg.best	barken
COR.58005.114.01	bark	yderste lag på et træ	sb.fk.sg.ubest.gen	barks
COR.58005.115.01	bark	yderste lag på et træ	sb.fk.sg.best.gen	barkens
COR.59594.110.01	bark	sejlskib	sb.fk.sg.ubest	bark
COR.59594.111.01	bark	sejlskib	sb.fk.sg.best	barken
COR.59594.112.01	bark	sejlskib	sb.fk.pl.ubest	barker
COR.59594.113.01	bark	sejlskib	sb.fk.pl.best	barkerne
COR.59594.114.01	bark	sejlskib	sb.fk.sg.ubest.gen	barks
COR.59594.115.01	bark	sejlskib	sb.fk.sg.best.gen	barkens
COR.59594.116.01	bark	sejlskib	sb.fk.pl.ubest.gen	barkers
COR.59594.117.01	bark	sejlskib	sb.fk.pl.best.gen	barkernes

Som et eksempel på den ortografiske variation kan vi se på COR-lemmaet *coronavirus*; det kan i ubestemt singularis staves både *coronavirus* og *koronavirus*; de tildeles hhv. 01 og 02 (se tabel 2).

TABEL 2. lemmaet 'coronavirus' i ubestemt singularis.

COR-id	lemma	glosse	gram. funk.	form
COR.53473.110.01	coronavirus		sb.fk.sg.ubest	coronavirus
COR.53473.110.02	coronavirus		sb.fk.sg.ubest	koronavirus

Hver af disse to stavemåder har tre forskellige flertalsformer, se tabel 3 nedenfor:

TABEL 3. lemmaet 'coronavirus' i ubestemt pluralis.

COR-id	lemma	glosse	gram. funk.	Form
COR.53473.112.01	coronavirus		sb.fk.pl.ubest	coronavira
COR.53473.112.02	coronavirus		sb.fk.pl.ubest	coronavirus
COR.53473.112.03	coronavirus		sb.fk.pl.ubest	coronavirusser
COR.53473.112.04	coronavirus		sb.fk.pl.ubest	koronavira
COR.53473.112.05	coronavirus		sb.fk.pl.ubest	koronavirus
COR.53473.112.06	coronavirus		sb.fk.pl.ubest	koronavirusser

Vi forbinder altså ikke de enkelte ortografiske variationer til hinanden, men nummererer dem sekventielt for hver grammatisk funktion.

Hvis man henter COR, vil man også lægge mærke til et normeringsfelt (som ikke er vist i eksemplerne i denne artikel): Det er normalt 1 (= *normeret*), men er nogle gange 0 (= *ikke normeret*); det drejer sig om enkelte bøjningsformer som er blevet autogenereret og derfor ikke bør bruges til fx stavekontrol.

5. Fra form til lemma

Informationerne i COR kan præsenteres på mange måder. På ordregister.dk kan man søge på COR-id, lemma og fuldform. Hvis man bruger den sidste mulighed, kan COR bruges til at finde ud af hvilke lemmaer og grammatiske funktioner en ordform kan referere til. Som et eksempel kan vi søge på ordformen *los*; resultatet heraf kan ses i tabel 4.

TABEL 4. ordformen 'los' i COR.

COR-id	lemma	glosse	gram. funk.	form
COR.22838.300.01	los		adj.sg.ubest.fk	los
COR.37712.209.01	losse		vb.imp	los
COR.55736.110.01	los	et dyr	sb.fk.sg.ubest	los
COR.70821.120.01	los	et spark	sb.itk.sg.ubest	los
COR.70821.122.01	los	et spark	sb.itk.pl.ubest	los
COR.78188.114.01	lo		sb.fk.sg.ubest.gen	los

6. Tre niveauer

Det Centrale Ordregister er opdelt i tre niveauer. Niveau 1 er det samme som grundregisteret og rummer altså de samme lemmaer som *Retskrivningsordbogen*.

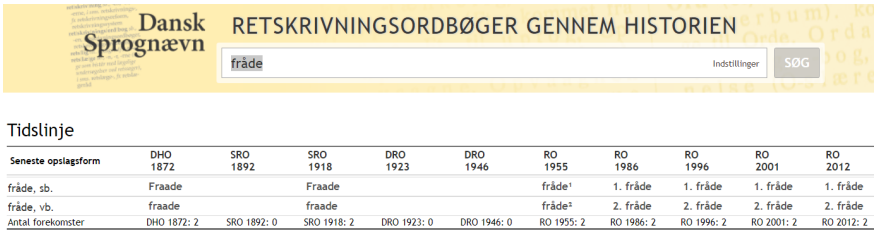
Niveau 2 rummer andre resurser fra de professionelle sprogmiljøer i Danmark (specifik dem som er medlem af Dansk Sprognævns repræsentantskab), og her vil mange andre nyttige resurser tilføjes med tiden. Allerede nu ligger her bl.a. en del ekstra lemmaer fra Den Danske Ordbog (udgivet af Det Danske Sprog- og Litteraturselskab [DSL]); denne resurse kaldes *COR.EXT*. Niveau 2 vil også komme til at rumme en semantisk udvidelse til grundregisteret produceret af DSL og CST (Center for Sprogteknologi ved Københavns Universitet). Se fx Nimb et al. (2022) for mere information om deres arbejde med at udvikle denne semantiske komponent.

Niveau 3 vil rumme alle andre resurser, og her er der ingen begrænsninger – ethvert relevant projekt kan få allokeret et præfiks og et id-interval hvis man henvender sig til Dansk Sprognævn.

Resurser kan relateres til andre resurser på samme eller lavere niveau. Alle resurser kan derfor markere relationer til grundresursen, men grundresursen hviler i sig selv og refererer aldrig til andre.

7. RO^{hist}

Dansk Sprognævns RO^{hist}-projekt (rohist.dk) er en søgemaskine hvormed man kan sammenligne de danske retskrivningsordbøger fra 1872 til 2012:



Seneste opslagsform	DHO 1872	SRO 1892	SRO 1918	DRO 1923	DRO 1946	RO 1955	RO 1986	RO 1996	RO 2001	RO 2012
fråde, sb.	Fraade		Fraade			fråde[1]	1. fråde	1. fråde	1. fråde	1. fråde
fråde, vb.	fraade		fraade			fråde[2]	2. fråde	2. fråde	2. fråde	2. fråde
Antal forekomster	DHO 1872: 2	SRO 1892: 0	SRO 1918: 2	DRO 1923: 0	DRO 1946: 0	RO 1955: 2	RO 1986: 2	RO 1996: 2	RO 2001: 2	RO 2012: 2

FIGUR 1. Et eksempel på en søgning i RO^{hist}.

Det er en videreudvikling af det svenske SAOLhist-system (som tillader parallelle opslag i *Svenska Akademiens ordlista över svenska språ-*

ket (1874, 1889-), *Ordbok öfver svenska språket* af A. F. Dalin (1850–1853) og Svenska Akademiens *Svensk ordbok* (2009). Se Diderichsen et al. (2015) for flere detaljer om sammenhængen mellem SAOLhist og RO^{hist}.

Vi arbejder løbende på at udvide RO^{hist} med alle historiske retskrivningsordbøger og andre ortografiske "rettesnore". Fx er vi for tiden ved at omdanne Ove Mallings *Store og gode Handlinger af Danske, Norske og Holstenere* fra 1777 til en ordbog som kan tilføjes til RO^{hist} (jf. Hartling & Widmann 2020).

Id-numrene i COR svarer til den nyeste udgave af *Retskrivningsordbogen* (p.t. 4. udgave fra 2012), men vi håber på at vi på sigt også kan tildele COR-numre til de historiske ordbøger i RO^{hist}. Disse ordbøger vil komme til at være niveau 2-resurser, og de vil derfor få deres eget præfiks og id-nummer-interval.

Man vil dog genbruge det samme id-nummer hvis lemmaet er det samme – også hvis stavemåden har ændret sig. Et ord som *fråse* (RO2012) vil altså få samme COR-nummer som *frådse* i RO1996 (*Retskrivningsordbogen* fra 1996) og som *fraadse* i DHO1872 (*Dansk Haandordbog* fra 1872):

COR-id	lemma	gram. funk.	form
COR.37337.200.01	fråse	vb.inf.akt	fråse
COR.RO2001.37337.200.01	fråse	vb.inf.akt	fråse
COR.RO1996.37337.200.01	frådse	vb.inf.akt	frådse
COR.DHO1872.37337.200.01	fraadse	vb.inf.akt	fraadse

Dette vil simplificere implementeringen af RO^{hist} betydeligt. Hvis man søger efter *fråse* i den fremtidige RO^{hist}, vil man blot skulle finde COR-nummeret (her 37337) og så ved et simpelt opslag konstatere hvorvidt dette er defineret i de historiske ordbøger.

De eksisterende links mellem ordbøgerne i RO^{hist} vil danne basis for dette arbejde. Vi vil altså tage den relationelle database som ligger til grund for RO^{hist}, analysere data og herudfra tildele historiske COR-idnumre. Det betyder også at hvis man finder en fejl i RO^{hist} – fx at en historisk stavemåde er blevet knyttet til det forkerte lemma – vil man blot skulle rette COR-id-nummeret i den historiske ordbog hvor fejlen er.

8. Uoverensstemmelser

Retskrivningsordbogen, og dermed COR's grundregister, rummer kun grundordforrådet, og når man indkoder en historisk ordbog, vil man derfor nogle gange stå med et lemma som ikke findes i den ældre ordbog. Det er dog ikke noget problem da alle ordbøger (og alle andre resurser som tilføjes til COR) vil få tildelt deres egne nummerområder til ekstra lemmaer.

Ordet *backfisch* var fx sidste gang med i 2001; hvis vi forestiller os at RO2001's ekstra nummerområde er 4002000–4004999, kan man så ved COR'ificeringen af denne ordbog tildele *backfisch* id-nummeret `COR.RO2001.4002123`.

Andre ordbøger kan så genbruge dette – *Backfisch* i DRO1946 (*Dansk Retskrivningsordbog* fra 1946) vil altså kunne hedde `COR.DRO1946.4002123`. Nummeret viser at det oprindeligt blev defineret af RO2001 ved at ligge i intervallet 4002000–4004999.

Det er heller ikke noget problem hvis stavemåder er blevet slået sammen eller splittet op i forhold til grundresursen eller de andre historiske ordbøger. COR gør det på niveau 2 og 3 muligt at bruge et såkaldt relationsfelt til dette formål. Dette er et ekstra felt i COR som indeholder et link til en eller flere andre COR-indgange, samt en type. Da grundresursen aldrig peger på andre resurser, bruges relationsfeltet aldrig her, men andre resurser kan bruge det til at vise hvordan afvigelser skal forstås. Forskellige resurser kan definere deres egne relationstyper afhængigt af deres behov; i arbejdet med de historiske retskrivningsordbøger regner vi med at få brug for flg.:

forkortelse	betydning
fus	fusion af to eller flere COR-indekser
rep	erstattet af et eller flere COR-indekser
spl	splittet op i to eller flere COR-indekser
sms	sammensætning af (til sammensatte ord)

Eksempelvis var der i DRO1923 (*Dansk Retskrivningsordbog* fra 1923) fri variation mellem formerne *Fjeder* og *Fjer* uafhængigt af betydningen, hvorimod vi i dag har to separate lemmaer. Det løser vi ved at lave et nyt id-nummer (her 4008020) som rummer information om at det har en relation til de to moderne indgange; se tabel 5.

TABEL 5. 'fjer' og 'fjeder'.

COR-id	lemma	glosse	gram. funk.	form	relation
COR.70131.110.01	fjeder		sb.fk.sg.ubest	fjeder	
COR.70759.110.01	fjer		sb.fk.sg.ubest	fjer	
COR.DRO1923.4008020.110.01	Fjeder		sb.fk.sg.ubest	Fjeder	fus:70759+70131
COR.DRO1923.4008020.110.02	Fjeder		sb.fk.sg.ubest	Fjer	fus:70759+70131
COR.DRO1923.70759					rep:4008020
COR.DRO1923.70131					rep:4008020

Det samme gør sig gældende hvis to historiske lemmaer er smeltet sammen til ét moderne – som et eksempel ses i tabel 6 lemmaet/lemmaerne *skade*; bemærk at de to lemmaer i RO1955 har samme indhold i relationsfeltet.

TABEL 6. 'skade'.

COR-id	lemma	glosse	gram. funk.	form	relation
COR.45662.110.01	skade	en fugl; en fisk	sb.fk.sg.ubest	skade	
COR.RO1955.4011080.110.01	skade	en fugl	sb.fk.sg.ubest	skade	rep:45662
COR.RO1955.4011081.110.01	skade	en fisk	sb.fk.sg.ubest	skade	rep:45662
COR.RO1955.45662					spl:4011080+4011081

(Fuglen *skade* har det latinske navn *Pica pica*; fisken *Dipturus batis*.)

RO^{hist} vil altså også skulle tjekke relationsfeltet for at kunne præsentere fyldestgørende resultater.

Planen er at opmærke alle de historiske ordbøger i omvendt kronologisk rækkefølge, altså i første omgang RO2001, og vores liste over kendte lemmaer vil derved gradvist vokse.

9. Næste udgave af *Retskrivningsordbogen*

Vi regner med at den næste udgave af *Retskrivningsordbogen* vil udkomme i 2024. I den forbindelse vil vi dels opdatere COR-grundserien så den fort-

sat afspejler nyeste retskrivning, dels publicere ændringerne mellem den gamle og den nye udgave i et maskinlæsbart format. Helt konkret vil den nuværende udgave få tildelt et nyt præfiks, COR.RO2012, og COR vil blive forbeholdt den nye udgave. Lad os rent hypotetisk forestille os at *sprog* ændres til *språg* i overensstemmelse med udtalen (men dette kommer helt sikkert ikke til at ske i virkeligheden!). COR-nummeret ændres ikke – det vil forsat være COR.40015.

Det er de konkrete former som vil blive ændret. De former vi har i dag, kan ses i tabel 7; efter ændringen vil de se ud som i tabel 8 nedenfor.

TABEL 7. lemmaet 'sprog' i COR.

COR-id	lemma	gram. funk.	form
COR.40015.120.01	sprog	sb.itk.sg.ubest	sprog
COR.40015.121.01	sprog	sb.itk.sg.best	sproget
COR.40015.122.01	sprog	sb.itk.pl.ubest	sprog
COR.40015.123.01	sprog	sb.itk.pl.best	sprogene
COR.40015.124.01	sprog	sb.itk.sg.ubest.gen	sprogs
COR.40015.125.01	sprog	sb.itk.sg.best.gen	sprogets
COR.40015.126.01	sprog	sb.itk.pl.ubest.gen	sprogs
COR.40015.127.01	sprog	sb.itk.pl.best.gen	sprogenes
COR.40015.129.01	sprog	sb.itk.sms	sprog-

TABEL 8. lemmaet 'sprog' i en hypotetisk fremtidig udgave af COR.

COR-id	lemma	gram. funk.	form
COR.40015.120.01	språg	sb.itk.sg.ubest	språg
COR.40015.121.01	språg	sb.itk.sg.best	språget
COR.40015.122.01	språg	sb.itk.pl.ubest	språg
COR.40015.123.01	språg	sb.itk.pl.best	språgene
COR.40015.124.01	språg	sb.itk.sg.ubest.gen	språgs
COR.40015.125.01	språg	sb.itk.sg.best.gen	språgets
COR.40015.126.01	språg	sb.itk.pl.ubest.gen	språgs
COR.40015.127.01	språg	sb.itk.pl.best.gen	språgenes
COR.40015.129.01	språg	sb.itk.sms	språg-

Og de gamle former vil nu få et RO2012-præfiks, så RO2012 dermed nærmest automatisk bliver tilføjet til de historiske ordbøger som kan tilgås i RO^{hist}, se tabel 9.

TABEL 9. lemmaet 'sprog' fra 2012-udgaven af Retskrivningsordbogen som det vil se ud når retskrivningen har ændret sig.

COR-id	lemma	gram. funk.	Form
COR.RO2012.40015.120.01	sprog	sb.itk.sg.ubest	sprog
COR.RO2012.40015.121.01	sprog	sb.itk.sg.best	sproget
COR.RO2012.40015.122.01	sprog	sb.itk.pl.ubest	sprog
COR.RO2012.40015.123.01	sprog	sb.itk.pl.best	sprogene
COR.RO2012.40015.124.01	sprog	sb.itk.sg.ubest.gen	sprogs
COR.RO2012.40015.125.01	sprog	sb.itk.sg.best.gen	sprogets
COR.RO2012.40015.126.01	sprog	sb.itk.pl.ubest.gen	sprogs
COR.RO2012.40015.127.01	sprog	sb.itk.pl.best.gen	sprogenes
COR.RO2012.40015.129.01	sprog	sb.itk.sms	sprog-

Denne information kan man så bruge til at generere en ændringsliste med; denne viser hvordan RO2012 forholder sig til denne hypotetiske fremtidige retskrivning i et format som kan bruges til fx at ændre en tekst med:

```
COR.RO2012.40015.120.01: sprog < språg
COR.RO2012.40015.121.01: sproget < språget
COR.RO2012.40015.122.01: sprog < språg
COR.RO2012.40015.123.01: sprogene < språgene
COR.RO2012.40015.124.01: sprogs < språgs
COR.RO2012.40015.125.01: sprogets < språgets
COR.RO2012.40015.126.01: sprogs < språgs
COR.RO2012.40015.127.01: sprogenes < språgenes
COR.RO2012.40015.129.01: sprog- < språg-
```

10. Opdatering af tekster og ordbøger til en anden udgave af retskrivningen

COR-opmærkning af historiske ordbøger har også mange andre leksikografiske anvendelser.

Man vil fx kunne ændre stavemåden af en COR-opmærket tekst fra en ortografi til en tidligere eller senere. Her er fx et fragment fra Mallings lærebog (1777):

… gandske forskiellige i Sprog, Sæder og Levemaade …

Hvis vi nu har en passende ordbog over Mallings sprog, gerne med fuldformer, kan vi nu tildele ordene `COR.MALLING`-id-numre. Disse kan vi så slå op i det moderne COR-register og derved opnå de moderne staveformer:

… ganske forskellige i sprog, sæder og levemåde …

I tabelform:

Malling	COR	RO2012
gandske	`COR.MALLING.10069.900`	ganske
forskiellige	`COR.MALLING.15189.303`	forskellige
i	`COR.MALLING.00852.880`	i
Sprog	`COR.MALLING.40015.122`	sprog
Sæder	`COR.MALLING.92121.112`	sæder
og	`COR.MALLING.00099.970`	og
Levemaade	`COR.MALLING.84179.110`	levemåde

Denne proces kan naturligvis bruges i begge retninger – man kunne lige så godt bruge den til at gøre en moderne tekst kunstigt gammeldags med.

En tilsvarende procedure kan anvendes på bl.a. ordbøger. Lad os eksempelvis antage at vi har en dansk-engelsk ordbog med COR-opmærkning af opslagsordene:

```
<entry>
  <orth COR="37337">frådse</orth>
  <pos>vb</pos>
  <tran>gorge</tran>
</entry>
```

Det vil nu være ganske let at slå op i COR og se at den gældende stavemåde er *fråse*. Man vil endda kunne gøre det næsten fuldautomatisk – der vil kun være behov for menneskelig assistance hvis en stavemåde splittes op i to, afhængigt af betydningen. Man kunne også bruge COR til at tilføje bøjningsoplysninger med.

11. COR-linkere

Til brug for korpuslingvistik og andre datalingvistiske anvendelser vil der også blive udviklet COR-linkere, dvs. programmer som tildeler det korrekte COR-id (med under-id, altså grammatisk kode) til alle ord i en løbende tekst. Når en tekst på denne måde er blevet COR-linket, vil det være trivielt at generere en ordklasseopmærket tekst (da alle de nødvendige informationer ligger i under-id'et), og hvis man har en COR-opmærket udtaleordbog, vil man også kunne tilknytte udtaleangivelser til alle ord (hvor også homografer får tilknyttet den korrekte udtale). Se Peter Juel Henrichsen (2023) for flere oplysninger om hvordan en sådan COR-linker kan se ud.

12. Norsk

Lad os til sidst til perspektivering se på muligheden for engang at forbinde dansk og norsk bokmål i værktøjer som RO^{hist}.

De to sprog deler som bekendt et ortografisk udgangspunkt, så når vi opmærker de historiske ordbøger med COR-id-numre, vil det blot kræve et norsk parallelprojekt for at kunne forbinde de to sprogs retskrivninger. Dette parallelprojekt kunne fx være en udvidelse af Metaordboka (Grønvik & Ore 2018; Ore et al. 2023).

Man kunne fx tage et dansk lemma som *sprog* og slå det op i COR-registeret; dets id-nummer er COR.40015. Vi kan så tage en resurse som er gammel nok til at den er en del af både dansk og norsk ortografihistorie, fx Ove Mallings læsebog *Store og gode Handlinger af Danske, Norske og Holstenere* (Malling 1777), som er ved at blive lavet om til en ordbog med COR-id'er (se Hartling og Widmann 2020), og tjekke at lemmaet også er defineret der: COR.MALLING.40015 *Sprog*. Hvis vi nu i fremtiden er så heldige at det hypotetiske norske parallelprojekt også har indekseret Malling-ordbogen, kunne det måske have id-nummeret `SOR.MALLING.123456` dér. Dette kan så bruges til at finde den moderne norske form med, her altså SOR.123456 *språk*.

Resultatet bliver ikke en tosproget ordbog, men forbindelser mellem de ord som har samme ortografiske ophav. Norsk *kveld* vil altså knyttes sammen med dansk *kvæld*, ikke med *aften*.

I teorien kunne man også lave forbindelser til andre sprog, men det kræver at nogen skaber disse links manuelt når man ikke kan gå direkte tilbage til sidste fælles retskrivning.

13. Konklusion

I denne artikel har vi givet en introduktion til Det Centrale Ordregister (COR) som tildeler unikke id-numre til alle lemmaer og ordformer på dansk, og vi har beskrevet grundregisterets struktur demonstreret hvordan nye resurser kan tilføjes. Vi har også set på forskellige leksikografiske anvendelser af COR, herunder RO^{hist}, som tillader sammenligning af forskellige historiske retskrivningsordbøger, og COR-linkere, som kan tildele det korrekte COR-id til alle ord i en løbende tekst. Endelig har vi diskuteret muligheden for at sammenknytte det danske COR med et hypotetisk norsk parallelprojekt.

Det Centrale Ordregister er frit tilgængeligt i dag på ordregister.dk. Vi håber at mange vil tilføje deres egne resurser til COR, så dansk dermed bliver et af de sprog som har de bedste resurser til sprogteknologiske og leksikografiske projekter.

Litteratur

DHO1872 = Grundtvig, Sven 1872. *Dansk Haandordbog med den af Kultusministeriet anbefalede Retskrivning*. 2. udgave. København: C. A. Reitzel.

Diderichsen, Philip, Anna Sofie Hartling, Anne Kjærgaard & Anna Kristiansen 2015. Retskrivningsordbøger gennem historien – følg retskrivningens udvikling ord for ord på http://rohist.dsn.dk. I Hansen, Inger Schoonderbeek & Tina Thode Hougaard (red.). *15. Møde om Udforskningen af Dansk Sprog*. Aarhus: Aarhus Universitet, 117-126.

DRO1923 = Glahder, Jørgen 1923. *Dansk Retskrivningsordbog. Udgivet af Undervisningsministeriets Retskrivningsudvalg*. København: Gyldendal.

DRO1946 = Glahder, Jørgen 1946. *Dansk Retskrivningsordbog. Udgivet af Undervisningsministeriets Retskrivningsudvalg. 5. Optryk*. København: Gyldendal.

Grønvik, O., & Ore, C.-E. S. 2018. Bokmål og nynorsk samindeksert – Metaordboka som verktøy for jamføring og utforsking av ordtilfang. *Nordiske studier i leksikografi 14*. Reykjavík, 87-95.

Hartling, Anna Sofie & Thomas Widmann 2020. Den første ortogra-

fiske rettesnor for dansk – fra læsebog til ordbog: Malling (1777) på http://rohist.dsn.dk. I: Goldshtein, Yonatan, Inger Schoonderbeek Hansen & Tina Thode Hougaard (red.). *18. Møde om Udforskningen af Dansk Sprog*. Aarhus: NORDISK, Institut for Kommunikation og Kultur, Aarhus Universitet, 213-230.

Henrichsen, Peter Juel 2023. Det Centrale Ordregister. Et indeks for det danske ordforråd – en gave til dansk sprogteknologi. I: Holmer, Louise et al. (red.), *Nordiska studier i lexikografi 16*. Lund & Göteborg: Nordiska föreningen för lexikografi, 113–126.

Malling, Ove 1777. Store og gode Handlinger af Danske, Norske og Holstenere.

Nimb, Sanni, Bolette S. Pedersen, Nathalie Carmen Hau Sørensen, Ida Flörke, Sussi Olsen & Thomas Troelsgård 2022. COR-S – Den semantiske del af Det Centrale OrdRegister (COR). *LexicoNordica* 29, 73-95.

Ore, Christian-Emil Smith, Oddrun Grønvik & Trond Minde 2023. Et fullformsystem for analyse av eldre tekst på tidlig nynorsk, bygd på Aasen-normalen. I: Holmer, Louise et al. (red.), *Nordiska studier i lexikografi 16*. Lund & Göteborg: Nordiska föreningen för lexikografi, 267–279.

RO1955 = *Retskrivningsordbog*. Dansk Sprognævn. 1955. København: Gyldendalske Boghandel.

RO1996 = *Retskrivningsordbogen*. 2. udgave. Dansk Sprognævn. 1996. København: Aschehoug.

RO2001 = *Retskrivningsordbogen*. 3. udgave. Dansk Sprognævn. 2001. København: Alinea – Aschehoug Dansk Forlag.

RO2012 = *Retskrivningsordbogen*. 4. udgave. Dansk Sprognævn. 2012. København: Alinea.