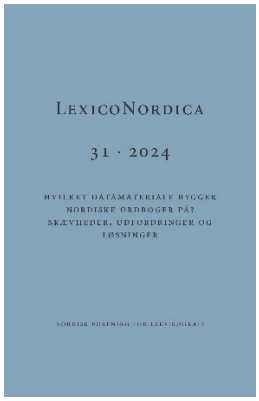
LexicoNordica

Titel:	Jagten på hverdagssproget – brugen af tekster fra internetfora i arbejdet med Den Danske Ordbog
Forfatter:	Kirsten Appel, Nathalie Hau Sørensen & Jonas Jensen
Kilde:	LexicoNordica 31, 2024, s. 39-59
URL:	https://tidsskrift.dk/lexn/issue/archive



Jagten på hverdagssproget – brugen af tekster fra internetfora i arbejdet med *Den Danske Ordbog*

Kirsten Appel, Nathalie Hau Sørensen & Jonas Jensen

This paper introduces methods and tools designed to alleviate the issues imposed by an increasingly uniform news wire corpus. The purpose is to help the editors of *The Danish Dictionary* (DDO) find common lemmas from everyday language which are underrepresented in the current corpus due to genre conventions and news values. The primary new tool, *Findor den yngre*, relies on a number of methods to filter the contents of a newly compiled corpus of chat room texts. The paper concludes that this complex filtering is superior to pure frequency-based methods when it comes to lemma selection, and that *Findor* and the new corpus provide editors with a more nuanced representation of contemporary Danish in general and a better selection of lemma candidates in particular.

1. Indledning

Den computerbaserede korpuslingvistiks fremkomst udgjorde et kvantespring i sprogforskningen og udarbejdelsen af ordbøger. Farvel til møjsommelig excerpering af belæg til alfabetisering i velordnede kartotekskasser, og goddag til søgbare databaser fulde af autentisk sprog, konkordanser der kunne sorteres efter ønske, og et overblik over sprogbrugen som man hidtil kun havde kunnet drømme om. Således var tidligere generationer af leksikografer lovligt undskyldt hvis de, grebet af tidsånden, antog at blot man havde et korpus, kunne man uden videre finde den objektive sandhed om sproget.

I tilfældet *Den Danske Ordbog* (DDO) havde man på sin vis mere at have det i da det redaktionelle arbejde indledtes i 1990'erne,

end det er tilfældet i dag: Ganske vist indeholdt DDO's oprindelige korpus blot 40 millioner ord, men det bød til gengæld på både tale- og skriftsprogsmateriale fra et bredt udvalg af kilder og genrer (dagbøger, nyhedsstof, skønlitteratur, transskriberet talesprog m.m.). Nutidens korpus er med sine 1,2 milliarder løbende ord en kæmpe i sammenligning, men hvad det har i størrelse, savner det i mangfoldighed.

Disse begrænsninger har givet anledning til at spørge om det overhovedet er muligt på grundlag af det nuværende korpus at "beskrive sproget sådan som det tales og skrives af et bredt udsnit af den danske befolkning" (*Fakta om DDO*), som var DDO's oprindelige målsætning. Det er desuden en overvejelse værd hvordan skævheden i korpus hidtil har påvirket DDO's lemmaselektion. Mest væsentligt i den aktuelle kontekst er imidlertid spørgsmålet: Hvordan kan en kombination af sprogteknologi og alternativer til det eksisterende korpus give redaktionen et mere nuanceret billede af "sproget sådan som det tales og skrives af et bredt udsnit af den danske befolkning"?

Denne artikels sigte er at besvare ovenstående spørgsmål gennem en præsentation og analyse af værktøjet *Findor den yngre*, der er udviklet netop med henblik på fremsøgning af lemmakandidater fra hverdagssproget i et til formålet kompileret korpus med tekster fra chatfora på internettet. For at forstå formålet med *Findor* er det nødvendigt at kende mere til dels det eksisterende korpus og dels baggrunden for første iteration af algoritmen, *Findor den ældre*. Begge dele beskrives i afsnit 2 nedenfor.

2. Baggrund

2.1. Korpus

Siden 2001 er det oprindelige, 40 millioner ord store DDO-korpus

løbende blevet udbygget med nye tekster, i altovervejende grad fra nyhedsmedier.

Resultatet er, som nævnt i afsnit 1, et korpus der er langt større end i den trykte ordbogs tid, men uden samme mangfoldighed. Det betyder at nogle sprogbrugere, fx børn og unge, er meget dårligt repræsenteret i vores korpus, mens professionelle sprogbrugere, især journalister, er kraftigt overrepræsenteret. Også korpussets ordvalg og emnemæssige sammensætning bærer præg af at teksterne er underlagt særlige krav til aktualitet og register. Trods adskillige metoder til fremfinding af lemmakandidater (bl.a. *Årets Ord*, *Månedens Ord*, brugerforslag, komposita og afledninger fra den trykte ordbog samt ord der optræder i redaktionel tekst) er det blevet tydeligt at der er ord og emner der er gået under radaren i den hidtidige lemmaselektion. Samtidig kan vi konstatere at ikke alt det vi faktisk finder, er lige velegnet til inklusion i DDO. Figur 1, som viser *Månedens Ord* fra 2023, illustrerer problemet, idet emner som krig og terror (fx *infanterikampkøretøj* og *terrorsag*) fylder uforholdsmæssigt meget, mens hverdagsemner (fx *læseglæde*) fylder tilsvarende lidt.

2.2. Automatiske metoder – hvad har andre prøvet?

I jagten på de tidligere oversete lemmakandidater i vores avistunge korpus blev den automatiske metode *Findor den ældre* (Sørensen et al. 2023) udviklet i 2023. Metoden viste at man med de rette sprogteknologiske værktøjer kan finde gode lemmakandidater blandt de lav- og mellemfrekvente ord i et aviskorpus.

Inspirationen til *Findor den ældre* kommer fra automatisk detektion af neologismer (Kerremans, Stegmayr & Schmid 2012, Falk, Bernhard & Gérard 2014, Langemets et al. 2020, Halskov & Jarvad 2010). Målet – at fremfinde henholdsvis lemmakandidater og neologismer – er omtrent det samme, dog med den væsentlige forskel at detektion af neologismer opererer med strengere kriterier for

Figur 1: Ord der optrådte på den frekvensbaserede liste over *Månedens Ord* i løbet af 2023. *Månedens Ord* udtrækkes ved at måle overhyppighed sammenlignet med andre måneder. Ordet *koranafbrænding* optræder tre gange, *koranlov*, *læseglæde* og *mobilitetsplan* optræder hver to gange. Resten af ordene optræder en enkelt gang.

hvad en god kandidat er. I stedet for kun at lede efter neologismer søger vi bredere efter ord som ikke nødvendigvis er nye i sproget, men som endnu ikke er opdaget af de frekvensbaserede værktøjer.

En bredt anvendt metode til automatisk detektion af neologismer er brugen af eksklusionslister. De bruges til at frasortere ord som man på forhånd ved ikke har interesse – fx ord som allerede findes i ordbogen, navnelister og stavefejl. Efter frasorteringen ved hjælp af eksklusionslisterne kan man så filtrere og postprocessere de resterende data til man får en mere overskuelig kandidatliste. Den største udfordring ved denne metode er effektiv frasortering

af støj i data som især kommer fra proprier, stavefejl, tokeniseringsfejl og gennemskuelige komposita.

I *Findor den ældre* omgås støjen gennem en lemmascore. Lemmascoren kombinerer information fra en række nøje konstruerede karaktertræk, også kaldet features, som er relevante for lemmaselektionen. Metoden kommer fra maskinlæringens feature engineering, hvor man bruger ekspertviden til at udvælge og udtrække en række egenskaber eller karakteristika om et fænomen som man vil modellere. Normalt vil man bruge features til at lave et datasæt man kan træne en model på. Denne metode er fx benyttet i Falk, Bernhard & Gérard (2014). Det kræver dog de rigtige data at træne en pålidelig model. Vi har positive eksempler fra online DDO, men mangler et tilstrækkeligt antal gode negative eksempler, det vil sige eksempler på ord der ikke er relevante for ordbogen (fx propriet *Juventus* eller den gennemskuelige sammensætning *hestenavn*). At en type ikke er med i DDO nu, er ikke et godt nok kriterium, da typen kan være en overset lemmakandidat. *Findor den ældre* brugte en mere simpel metode, nemlig et vægtet gennemsnit med nøje manuelt justerede vægte, som viste sig at være tilstrækkelig effektiv. Vi har nu videreudviklet *Findor* og præsenterer i denne artikel *Findor den yngre* som i stedet for avistekster er målrettet chatforumtekster.

3. Metode

3.1. Indsamling af internettekster

Som nævnt i afsnit 2.1 udgør den store overvægt af avistekster en udfordring ved vores nuværende korpus, og vi har derfor sat os for at finde tekster der både er frit tilgængelige, og som kan udgøre en modvægt til det eksisterende korpus.

Vi mangler først og fremmest hverdagssprog – ting vi taler om

sammen, men som ikke nødvendigvis lever op til nyhedskriterierne, og som derfor sjældent kan genfindes i avisartikler. En oplagt kilde til dette sprog er åbne internetfora hvor der typisk skrives i et uformelt sprog om hverdagsproblemstillinger, fx hvordan man undgår at blive syg i vinterperioden, eller hvilken tørretumbler andre kan anbefale.

Sproget på internetfora adskiller sig fra avisteksternes på flere områder. For det første er brugerne ikke professionelle skribenter, og teksterne er i mindre omfang redigeret og korrekturlæst, og vi må derfor forvente at se flere stavefejl. Teksterne er desuden kortere, mere uformelle, og de lægger op til interaktion. Vi ser fx en del emojis og slang i data. Teksterne bærer præg af at handle om hvad skribenten har på hjerte nu og her, altså øjeblikstanker, i modsætning til avisteksternes krav om relevans for en bred læserskare.

Da internetfora ligger frit tilgængelige på internettet, kan data derfra høstes via web scraping (det vil sige automatisk indsamling af internettekster). Vi har identificeret syv forskellige internetfora med hverdagssprog (Hestenettet, baby.dk, bold.dk, HardwareOnline, Pokernet, debatten.net og reddit.com/r/denmark). Hvert forum indeholder tekster fra perioden 2005-2023. Derudover har vi udvalgt specifikke subfora med en forventning om at de indeholder meget hverdagssprog. Disse subfora har typisk overskrifter som "fri-snak-fredag", "hyggesnak", "off-topic", "generelt" osv. Vi har bestræbt os på at indhente tekst fra fora med forskellige emneområder, fx heste, graviditet, sport og teknik. Til hvert forum har vi skræddersyet et pythonscript der gemmer indlæg og kommentarer som separate tekstfiler med følgende metadata: id, tekstens titel, selve teksten, sektion, url, type, dato, parent id og parent url. De to sidstnævnte sikrer at vi kan genskabe indlæggets kontekst. I alt har det resulteret i et korpus på 174 millioner løbende ord.

3.2. Værktøjet *Findor den yngre*

Findor den yngre bygger på samme arkitektur som sin forgænger, *Findor den ældre*, nemlig de tre faser: 1) præprocessering af korpus, 2) beregning af en score for "lemmahed" (altså egnethed som lemma i DDO) og 3) sortering efter denne score. Forskellen på de to iterationer af *Findor* er datagrundlaget (avistekster versus chatforumtekster) og indmaden i fase 1 og 2. Vi vil i det følgende kun gå i dybden med opbygningen af *Findor den yngre*, og betegnelsen *Findor* vil derfor referere til den nyeste version. For en grundig beskrivelse af *Findor den ældre* henviser vi til Sørensen et al. (2023).

For at kunne undersøge kvaliteten af *Findor* deler vi DDO's materiale op i det der optrådte i den trykte udgave (2002-2005), og det som siden er blevet føjet til onlineudgaven (2006-). I opbygningen af *Findor* bruger vi kun information tilgængelig i den trykte udgave og gemmer derfor alle opdateringer som en "guldstandard" der kan bruges som led i evalueringen.

3.2.1. Fase 1: Præprocessering

Formålet med præprocesseringen er at gå fra korpus i rå tekst til en bruttoliste med lemmakandidater. Det vil sige at ikke alle ord på listen nødvendigvis er gode lemmakandidater, men at vi sorterer alt fra som vi allerede ved ikke har interesse for DDO. Da det kan diskuteres om alt indhold på listen kan kaldes ord eller lemmaer, vil vi bruge "tokens" om enkelte eksempler på vilkårlige tekststrenge, "typer" om en samlet gruppe af tokens med samme form, og "kandidater" om typer som er behandlet af *Findor*.

Præprocesseringen tager udgangspunkt i et årsopdelt korpus (årgangene 2005-2023). Herefter udfører vi en simpel tokenisering ved at adskille teksten med mellemrum. Vi renser også data yderligere ved at fjerne URI'er, transformere store bogstaver til små og ved at fjerne tokens som indeholder ugyldige tegn (fx ه¤þü)[1]. Vi

1 Gyldige tegn = abcdeéfghijklmnopqrstuvwxyzæøå0123456789-&/.´"¹²³⁴⁵⁶⁷⁸⁹⁰₀₁₂₃₄₅₆₇₈₉ₓ

fjerner også al tegnsætning undtagen bindestreg, som bruges til at rense tokeniseringsfejl. Herefter grupperer vi tokens som typer og optæller deres frekvens for hvert år og i hele korpusset. Frekvensoptællingerne tillader os at fjerne alle typer som har en frekvens på mindre end fem. Denne liste med typer og frekvenser udgør den første kandidatliste.

På dette stadie indeholder kandidatlisten stadig uinteressante typer, det vil sige enten støj, navne eller ord som allerede er i den trykte ordbog, og som derfor ikke skal tilføjes til DDO. Formålet med det næste trin er derfor at fjerne så mange uinteressante typer som muligt. Vi fjerner derfor i første omgang alle typer som indeholder tal, eller som starter med en bindestreg. Dernæst fjerner vi alle typer som forekommer på én af fem forskellige eksklusionslister: en fuldformsliste fra den trykte ordbog og fire lister med fornavne (mandlige og kvindelige), efternavne og stednavne som er registreret af Danmarks Statistik. Da flere typer kan være eksempler på det samme ord i forskellige bøjninger, bruger vi også værktøjet *CSTLEMMA* til automatisk at lemmatisere typerne og gruppere dem igen. Til sidst fjerner vi alle typer som optræder i færre end tre årgange. En oversigt over antallet af typer efter hvert rensningstrin kan ses i tabel 1:

Trin	Fjernet	Antal på kandidatliste
Første kandidatliste		335.401
Fjern tal og bindestreg	7.494	327.907
I den trykte DDO	122.482	205.425
Findes på navnelister	23.340	182.085
Lemmatisering	21.263	160.822
Mindre end tre årgange	25.491	135.331

Tabel 1: Antal typer der resterer efter de respektive trin i præprocesseringen.

3.2.2. Fase 2: Udregning af "lemmahed"

Kernen i *Findor* er en måling af "lemmahed", som vi kalder lemmascoren. Lemmascoren er et vægtet gennemsnit af flere delscorer som hver især afspejler et karaktertræk ('feature') der kan være relevant for lemmaselektionen. Det er blandt andet stabilitet over tid, overensstemmelse med dansk ortografi og morfologi samt semantisk lighed med lemmaer som allerede er i DDO. I denne fase bliver ingen typer fjernet på baggrund af et enkelt kriterium som i præprocesseringen. I stedet skaber det vægtede gennemsnit et samlet billede på tværs af delscorerne, så en dårlig score ét sted kan opvejes af flere gode scorer andre steder. I det følgende forklarer vi baggrunden for hver delscore, og hvordan den er udregnet.

Stabilitet over tid

Et kriterium i DDO's lemmaselektion er at et ord skal optræde stabilt i korpus hen over en årrække. Vi antager derfor at en mere udbredt tidsmæssig repræsentation i korpus korrelerer med egnethed som lemma, hvorfor vi tæller antallet af årgange et ord optræder i – uden dog at tage højde for frekvensen pr. år. Scoren for stabilitet over tid er antallet af år en type forekommer i, delt med det totale antal år i korpus (18). Dermed er scoren højere, desto flere år en type optræder i.

Bøjningsformer

Vi antager at et ord med større sandsynlighed er etableret i dansk hvis det følger genkendelige mønstre for dansk morfologi. Som led i den automatiske lemmatisering i præprocesseringen (se afsnit 3.2.1) har vi optalt og gemt antallet af unikke former et ord optræder i. Vi har ikke taget højde for ordklasse i optællingen. Derfor kan ordklasser med få mulige bøjningsformer (fx adverbier) risikere at blive nedprioriteret i denne score. Vi kan dog også se at det er få typer som har mere end tre former i vores data. Vi vurderer derfor at uligheden har begrænset indflydelse på den en-

delige lemmascore. Scoren for bøjningsformer er antallet af unikke former for en type, men justeret så scoren ligger mellem 0 og 1.

Frasortering af proprier
I præprocesseringen har vi allerede fjernet en lang række person- og stednavne fra kandidatlisten. Vi mangler dog stadig at frasortere andre proprier som ikke forekommer på vores eksklusionslister. Det gælder bl.a. firma- og produktnavne og personnavne fra andre kulturer. Til dette formål anvender vi en sprogteknologisk metode ved navn Named Entity Recognition igennem modellen *ScandiNER*. Named Entity Recognition går ud på at få en model til at genkende proprier i en kort tekst, typisk én sætning ad gangen. For hver type har vi derfor tilfældigt udvalgt op til ti sætninger fra korpus og tagget dem ved hjælp af modellen. Scoren er den procentvise andel af sætninger hvor en type blev tagget som et proprium. Denne score vægtes negativt, og dermed er scoren lavere, jo oftere en type bliver anset for at være et proprium.

Dansk ortografi og nedgradering af fremmedord
For at give ord der følger dansk ortografi, større vægt end fx tokeniseringsfejl og fejlstavninger, har vi udviklet en score som måler "danskhed", altså overensstemmelse med typisk dansk ortografi. Til formålet har vi trænet en tetragrammodel på en fuldformsliste fra DDO, som beregner sandsynligheden for at en bogstavsekvens ligner kendte danske ord. Scoren er enten 0 eller 1, alt efter om sandsynligheden overstiger en given grænseværdi.

Der kan også være fremmedord på listen som er relevante lemmakandidater, især anglicismer og indlån fra engelsk og tysk. Vi har derfor tilføjet endnu en score som ser på om en type følger dansk, engelsk eller tysk ortografi. Vi har derfor også trænet en tetragrammodel på ordlister for engelsk (*Moby Crosswords word list*) og tysk (*Aspell-de*). Sprogscoren er højest hvis en type med stor sandsynlighed følger dansk ortografi, næsthøjest hvis typen

følger engelsk ortografi, og tredjehøjest hvis den følger tysk ortografi. Sprogscoren er 0 hvis typen er under grænseværdien for alle sprogene.

Som noget nyt i forhold til *Findor den ældre* har vi indført endnu en score der ser på typernes kontekst. I stedet for kun at se på ordene isoleret undersøger vi nu også om en type optræder i en dansk kontekst eller en kontekst fra andre sprog. På denne måde kan vi adskille faktiske indlån, som ofte vil optræde blandt danske ord, fra kodeskift eller citater på et andet sprog, som netop ikke vil have mange danske ord omkring sig. Denne score er særligt relevant for internetdata, da vi ikke kan være sikre på at der altid skrives på dansk i de webscrapede tekster. Til at udregne denne score bruger vi de samme tilfældigt indhentede sætninger fra propriumscoren, og vi bruger vores tetragrammodeller til at finde det mest sandsynlige sprog for hvert token inden for fem pladser fra den undersøgte type. Indlånsscoren er den gennemsnitlige sprogscore for alle tokens i alle sætningerne. Det resulterer i en højere score hvis ordet optræder i danske kontekster.

Semantisk lighed

Vi antager at et ord er en bedre lemmakandidat hvis det har synonymer eller nærsynonymer der allerede optræder i DDO. Vi har derfor benyttet en word2vec-model til at finde de 20 nærmeste naboer for hvert ord på vores lemmaliste. Det giver en højere score hvis ordet har flere synonymer i ordbogen blandt sine nærmeste naboer. I Sørensen et al. (2023) anvendtes en word2vec-model for dansk trænet på avistekster, men til denne undersøgelse har vi trænet en ny model specifikt på internettekster. Denne foranstaltning sikrer at alle ord i materialet faktisk er repræsenteret i vores semantiske model.

4. Resultater

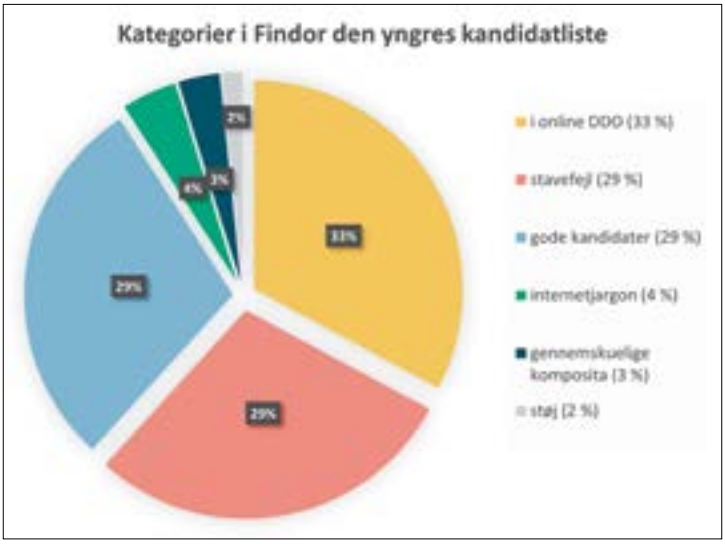
Vi har kørt *Findor* på de 174 millioner løbende ord i chatforumkorpusset, og det har resulteret i 135.331 unikke lemmakandidater rangeret efter lemmascore. Det er ikke tanken at alle kandidater skal med i DDO, men snarere at redaktørerne kan vælge fra et filtreret udsnit af kandidatlisten.

Vi evaluerer *Findors* kandidatliste på tre måder. Først analyserer vi indholdet i kandidatlistens øverste 2.000 kandidater. Herefter bruger vi DDO's opdateringer siden 2005 til at sammenligne *Findors* fundne lemmakandidater med lemmakandidater fundet ved ren frekvens. Til sidst foretager vi en kvalitativ evaluering, hvor en leksikograf manuelt vurderer tre forskellige udsnit af *Findors* kandidatliste.

4.1. Hvad finder *Findor*?

I figur 2 ses de øverste 2.000 typer på den kandidatliste som *Findor* har genereret, fordelt på seks forskellige kategorier. Kategorien "i online DDO" er automatisk annoteret og indeholder de lemmaer som er blevet tilføjet til DDO siden 2005. De resterende kategorier er annoteret manuelt. Med "gennemskuelige komposita" menes fx *laksetærte* og *broccolitærte* der ikke er decideret uegnede som lemmakandidater, men som det på grund af deres gennemskuelighed aktuelt er mindre oplagt at føje til ordbogen end mere svært afkodelige alternativer som *pletbløde*, *gummirøjser* og *pigefarve*. Sidstnævnte tilhører de "gode kandidater", hvilket vil sige at de er gode nok til at komme i betragtning til DDO.

Resultaterne er lovende da 62 % af top-2.000 enten allerede er i DDO eller egner sig til at komme det. Der er dog stadig en del stavefejl – mere herom i afsnit 5.3.



Figur 2: Kandidatlistens øverste 2.000 ord fordelt på kategorierne "i online DDO", "stavefejl", "gode kandidater", "internetjargon", "gennemskuelige komposita" og "støj".

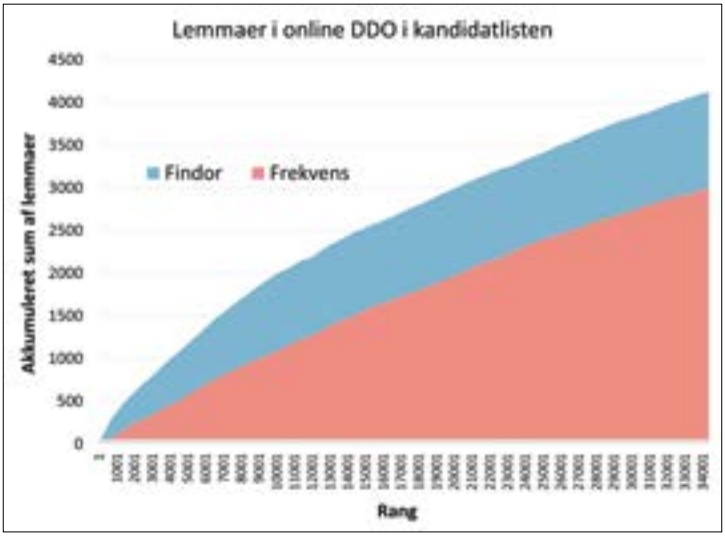
4.2. *Findor* versus frekvens

Findor den yngre er udviklet til at finde kandidater i nyindsamlet materiale som vi ikke tidligere har bearbejdet med vores gængse, frekvensbaserede værktøjer. Et oplagt spørgsmål er derfor om *Findor den yngre* klarer sig bedre end ren frekvens, når det gælder om at finde lemmakandidater i et helt nyt materiale, eller om vi kunne have opnået samme resultat med ren frekvens.

I denne undersøgelse bruger vi de lemmaer som er tilføjet til DDO siden 2005 som guldstandard, jf. afsnit 3.2. Vi antager at jo flere DDO-lemmaer en metode finder og placerer højt på listen, desto bedre er metoden til at finde nye lemmakandidater.

Vi sammenligner *Findors* rangering af lemmakandidaterne med en rangering fra højeste til laveste frekvens. I figur 3 ses den akkumulerede sum af lemmaer fra onlineudgaven af DDO for de

øverste 35.000 kandidater fra *Findor* (blå) og frekvens (rød). Her kan vi se at *Findor* konsekvent finder flere DDO-lemmaer, men er bedst i toppen af listen, hvor kurven er stejlest. Fx er 51 af *Findors* top-100 med i online DDO, svarende til 51 % af lemmaerne, mens det samme kun gør sig gældende for 12 lemmaer eller 12 % for frekvenslisten. Det samme mønster ser vi for top-1.000 hvor 37 % af *Findors* lemmaer er med i online DDO, mens kun 11 % er det for frekvenslisten.



Figur 3: Akkumuleret sum af antal lemmaer der allerede findes i onlineudgaven af DDO, fundet af henholdsvis *Findor* og via ren frekvens.

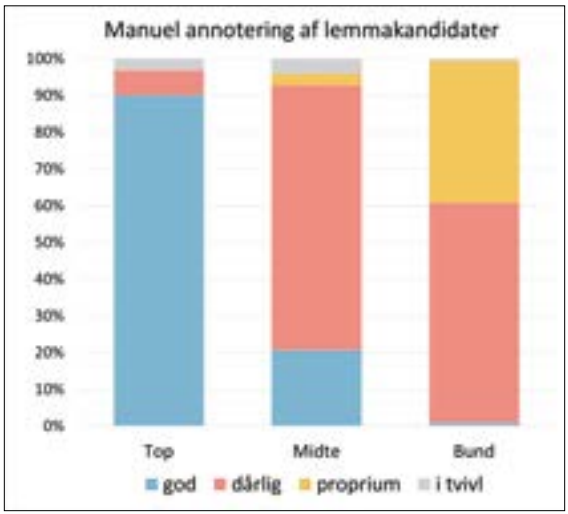
4.3. Manuel evaluering

En ulempe ved at bruge DDO's egne tilføjelser som guldstandard er at kandidaterne i listen sagtens kan være gode lemmakandidater uden at være føjet til ordbogen endnu. Dertil kommer at formålet med at udvikle *Findor* netop var at finde ord som vi ikke

tidligere har fundet – og som vi måske end ikke tidligere har haft adgang til at finde fordi vores sædvanlige korpus ikke indeholder de pågældende ord. For at kunne evaluere om metoden også fanger interessante kandidater uden for DDO's lemmaliste, iværksatte vi en evaluering med manuel annotation.

Vi udvalgte et tilfældigt udsnit på 222 kandidater fra henholdsvis de øverste, midterste og nederste 2.000 ord fra kandidatlisten, i alt 666 kandidater. Først frasorterede vi stavefejl, dernæst blandede vi de tre lister sammen, således at top-, midter- og bundkandidaterne optrådte sammen på en randomiseret liste.

Vi gav efterfølgende den samlede liste til en erfaren leksikograf med den opgave at placere hver kandidat i én af de fire kategorier "god", "dårlig", "proprium" og "i tvivl". Resultatet kan ses i figur 4. Her fremgår det at 90 % af ordene taget fra de øverste 2.000 er gode lemmakandidater, mens det samme kun gælder for 21 % fra midtergruppen og 1 % fra bunden.



Figur 4: Antal lemmaer vurderet af en leksikograf til at passe i kategorien "god", "dårlig", "proprium" eller "i tvivl" fra toppen, midten og bunden af listen over lemmakandidater.

Ud af de i alt 247 lemmaer som leksikografen har givet kategorien "god", findes 207 af dem ikke i forvejen på vores interne lister over lemmakandidater. *Findor* har altså overvejende fundet nye lemmakandidater.

Det er desuden værd at bemærke at proprierne helt overvejende befinder sig i den nederste del af listen, og at *Findor* altså har haft succes med at nedprioritere navne.

5. Diskussion

5.1. Repræsentativitet

Afsættet for denne undersøgelse er netop vores eksisterende korpus' repræsentativitet – eller mangel på samme. Vi er som nævnt i afsnit 2.1 blevet tiltagende bevidste om de mange domæner og sprogbrugere som ikke optræder i vores korpus, og forsøg på at udbygge det eksisterende korpus har endnu ikke båret frugt. Det er især juridiske udfordringer der spænder ben for denne proces, idet de mest attraktive teksttyper (fx moderne skønlitteratur) er belagt med copyright.

At sammensætte et repræsentativt korpus er af praktiske, økonomiske og ikke mindst juridiske årsager så godt som umuligt, men det er værd at huske på at en repræsentativ og fordomsfri beskrivelse af sproget er endnu mere umulig uden adgang til et korpus. Vi mener derfor ikke at manglen på et perfekt afbalanceret korpus skal spænde ben for at arbejde med et tekstkorpus, men vi må som leksikografer erkende korpussets fejl og mangler og ikke drage konklusioner ud over hvad vores korpus rent faktisk kan bære. Samtidig har vi et ansvar for gradvis at gøre vores korpus mere repræsentativt. Nærværende undersøgelse har netop til hensigt at råde bod på den aktuelle skævhed i forhold til genrer og sprogbrugere, idet vi har tilføjet tekster fra chatfora. Et internetkor-

pus er selvsagt ikke i sig selv mere repræsentativt for skriftsprog i Danmark end et aviskorpus, men ved at stykke forskellige korpusser sammen får vi et mere alsidigt og balanceret indtryk af moderne dansk, idet flere sprogbrugere og domæner er repræsenteret.

5.2. Substantiver og komposita

Et kritikpunkt ved *Findor den ældres* lemmakandidater var den relativt høje andel af såkaldt gennemskuelige komposita. Vi fjernede derfor værktøjets kompositumsplitter, som vi formodede kunne øge mængden af komposita, netop fordi den belønnede ord der var sammensat af eksisterende ord i DDO, med en højere score.

Findor den yngre finder fortsat først og fremmest komposita. DDO har imidlertid eksisteret i omkring 30 år, og vi må derfor også forvente at beskrivelsen af kerneordforrådet er et afsluttet kapitel. Vi er længere ude i periferien, og derfor er der nødvendigvis færre simpleksord – og i det omfang vi stadig tilføjer simpleksord til ordbogen, er de som oftest låneord. Vi vil på et senere tidspunkt undersøge om vi kan fremfinde de mest relevante sammensætninger for DDO, fx ved at kigge på brugsstatistikker som det er gjort på norsk i Paulsen (2023).

De lemmakandidater *Findor* har fundet, er i altovervejende grad substantiver. Det skyldes dels det faktum at netop denne ordklasse er meget produktiv (bl.a. i kraft af de mange komposita), dels at *Findors* parametre for søgning og filtrering især tilgodeser substantiver. Det kunne være interessant i fremtiden at forsøge målrettet at opspore eksempelvis adjektiver, men det har ikke været prioriteten i denne omgang.

5.3. Stavefejl

Et iøjnefaldende resultat er mængden af stavefejl på den endelige lemmakandidatliste. Vi forventede, jf. afsnit 3.1, flere stavefejl i et

chatforumkorpus sammenlignet med vores aviskorpus, fordi brugere af internetfora sjældent er professionelle sprogbrugere med adgang til korrekturlæsere og redaktører, men 29 % stavefejl oversteg alligevel vores forventning.

En mulig forklaring er vores brug af en word2vec-model. Modellen finder nemlig semantisk lignende ord, og stavefejl har selvsagt en betydning der er identisk med den korrekt stavede udgave. Denne forklaring underbygges af at stavefejlene i altoverskyggende grad befinder sig i toppen af listen og altså er blevet vægtet højt af *Findor*.

Imidlertid er mængden af stavefejl ikke udelukkende et problem for os: Dels føjer vi dem til vores liste over fejlstavninger, sådan at de udelades fra fremtidige lister med lemmakandidater, dels er stavefejl en ressource i sig selv, idet vi indlemmer dem i DDO's søgehjælp, sådan at brugere i fremtiden får nemmere ved at finde det ønskede opslagsord – uanset stavekundskaber.

5.4. Sorterer vi for meget fra?

At sortere store mængder støj fra er hele formålet med vores algoritme, men det er klart at gode lemmakandidater også kan ryge i svinget. Som nævnt i afsnit 3.2.1 frasorterer vi alle ord der indeholder tal, for fx at slippe for produktive dannelser som *1-1-sejr* – til gengæld ofrer vi muligheden for i denne omgang at finde gode lemmakandidater der indeholder tal. Det samme gør sig gældende for låneord, idet vi ved hjælp af en tetragrammodel nedprioriterer ord der ikke følger dansk, engelsk og tysk ortografi. Det betyder at vi måske går glip af gode kandidater fra andre sprog, eksempelvis *romanesco*. På samme måde frasorteres forkortelser som *uvb-stråle* 'ultraviolet stråle fra fx solen'. Det er dog ikke tanken at *Findors* lemmakandidatliste skal stå alene, og vi anser det derfor ikke som et problem at vi måske frasorterer for meget – for frasortering er en forudsætning for hele metoden.

6. Konklusion

Ved hjælp af *Findor den yngre* har vi fundet lemmakandidater fra hverdagssproget som vi ellers ikke kunne finde automatisk. *Findor* er rent frekvensbaserede værktøjer overlegen i kraft af sin komplekse filtrering der tager højde for bl.a. lemmakandidaternes semantiske lighed med eksisterende lemmaer og deres overensstemmelse med dansk ortografi og morfologi. Konkret udmærker 207 af de 247 ord som leksikografen godkendte fra *Findors* liste, sig ved at være nye i forhold til de kandidater vi har fundet via vores hidtidige, rent frekvensbaserede metoder.

Findor retter i et vist omfang op på aviskorpussets iboende begrænsninger – men metoden bør ad åre afprøves på og tilpasses et endnu bredere udvalg af teksttyper og -genrer.

Litteratur

Ordbøger, korpusser og digitale resurser

Aspell-de. <ftp.gnu.org/gnu/aspell/dict/oindex.html> (januar 2024).

Baby.dk. <baby.dk/debat/grupper.aspx> (december 2023).

Bold.dk. <bold.dk/snak> (november 2023).

CSTLEMMA. <github.com/kuhumcst/cstlemma> (oktober 2023).

Danmarks Statistik. <sprogteknologi.dk/dataset/fornavne-og-efternavne-i-befolkningen-i-danmark-i-januar-2020> (marts 2023).

DDO = *Den Danske Ordbog*. Det Danske Sprog- og Litteraturselskab. <ordnet.dk/ddo> (marts 2024).

Debatten.net. <debatten.net/forum/> (december 2023).

Fakta om DDO. <ordnet.dk/ddo/fakta-om-ddo/ordbogens-tilblivelse> (marts 2024).

HardwareOnline. <hardwareonline.dk/forum_list.aspx?fid=23> (november 2023).
Hestenettet. <heste-nettet.dk/forum/1/> (august 2023).
Moby Crosswords word list. <gutenberg.org/files/3201/files/> (januar 2024).
Pokernet. <pokernet.dk/forum/kategorier/frontpage/off-topic.html> (december 2023).
Reddit.com/r/denmark. <reddit.com/r/denmark> (november 2023).
ScandiNER. <huggingface.co/saattrupdan/nbailab-base-ner-scandi> (januar 2024).
Tetragrammodel for dansk. <github.com/dsldk/lexiscore> (januar 2024).
Word2vec-model for dansk. <korpus.dsl.dk/resources/details/word2vec.html> (oktober 2023).
Wordclouds. <wordclouds.com> (marts 2024).

Anden litteratur

Falk, Ingrid, Delphine Bernhard & Christophe Gérard (2014): From Non Word to New Word: Automatically Identifying Neologisms in French Newspapers. I: *LREC-The 9th edition of the Language Resources and Evaluation Conference*. Reykjavik, Iceland. 4338-4344. <lrec-conf.org/proceedings/lrec2014/pdf/288_Paper.pdf>.
Halskov, Jakob & Pia Jarvad (2010): Manuel og maskinel excerpering af neologismer. I: *NyS – Nydanske Sprogstudier* 38, 39-68.
Kerremans, Daphné, Susanne Stegmayr & Hans-Jörg Schmid (2012): The NeoCrawler: Identifying and Retrieving Neologisms from the Internet and Monitoring Ongoing Change. I: Kathryn Allan & Justyna A. Robinson (eds.): *Current Methods in Historical Semantics* 73. Berlin, Boston: De Gruyter Mouton. 59-96.

Langemets, Margit, Jelena Kallas, Kaisa Norak & Indrek Hein (2020): New Estonian Words and Senses: Detection and Description. I: *Dictionaries: Journal of the Dictionary Society of North America* 41(1), 69-82.

Norling-Christensen, Ole & Jørg Asmussen (1998): The Corpus of the Danish Dictionary. I: *Lexikos* 8, 223-242. doi.org/10.5788/8-1-955.

Paulsen, Mikkel Ekeland (2023): Wheat or Chaff? A Compound Selection Model Based on Look-Up Data. I: *International Journal of Lexicography* 36(3), 306-324.

Sørensen, Nathalie Hau, Nicolai Hartvig Sørensen, Kirsten Lundholm Appel & Sanni Nimb (2023): Trawling the corpus for the overlooked lemmas. I: Marek Medveď, Michal Měchura, Carole Tiberius, Iztok Kosem, Jelena Kallas, Miloš Jakubíček & Simon Krek (eds.): *Electronic lexicography in the 21st century (eLex 2023): Invisible Lexicography. Proceedings of the eLex 2023 conference*. Brno, 27-29 June 2023. Brno: Lexical Computing CZ s.r.o. 392-409.

Kirsten Appel
Seniorredaktør
Det Danske Sprog- og Litteraturselskab
Christians Brygge 1
DK-1219 København K
ka@dsl.dk

Jonas Jensen
Seniorredaktør
Det Danske Sprog- og Litteraturselskab
Christians Brygge 1
DK-1219 København K
jj@dsl.dk

Nathalie Hau Sørensen
Assisterende redaktør
Det Danske Sprog- og Litteraturselskab
Christians Brygge 1
DK-1219 København K
nats@dsl.dk