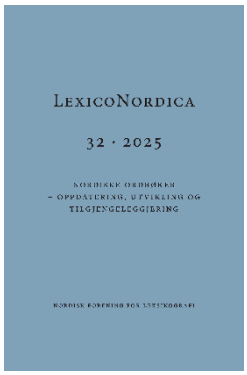
LexicoNordica

Titel:	Den universelle First-Letter Law og dens anvendelse til statistisk undersøgelse af skandinaviske ordbøger
Forfatter:	Poul Hansen
Kilde:	LexicoNordica 32, 2025, s. 243-260
URL:	https://tidsskrift.dk/lexn/issue/archive



Den universelle First-Letter Law og dens anvendelse til statistisk undersøgelse af skandinaviske ordbøger

Poul Hansen

The article demonstrates how the mathematically based method First-Letter Law can utilize the statistical properties of lemmalists to analyse specific issues in Nordic lexicography. This offers a significant advantage compared to traditional approaches.

The practical implementation on three Scandinavian dictionaries and selected sub-materials (Swedish prefixed verbs) includes: 1) a clear model-based description of the frequency and relationships of initial letters in a dictionary, 2) calculation of the correlation between two dictionary materials, and 3) statistical documentation of diachronic changes in dictionaries.

The article reviews various aspects of the method and its limitations. The reader also gets practical instructions for applying the First-Letter Law in Excel.

1. Indledning

Enhver leksikograf er bekendt med, at fordelingen af begyndelsesbogstaver i en lemmaliste er ganske ujævn. Det er også velkendt, at det handler om et konsekvent mønster, der ikke adskiller sig meget fra ordbog til ordbog. Vi ved i dag, at mønstret kan beskrives tilfredsstillende med en eksponentiel funktion. Et lignende fænomen ses også hos cifre, og her udviklede Newcomb (1881) og senere Benford (1938) tidligt en matematisk funktion, der beskriver fordelingen af førstecifre. I litteraturen kaldes denne funktion Benfords lov, Newcomb-Benfords lov eller BL. Med hensyn til tekst skete modeludviklingen lidt senere og over en længere periode, men i dag er der en række modeller, som beskriver Ben-

ford-lignende tendenser og andre fænomener inden for korpuslingvistikken, herunder Zipfs lov (Zipf 1936), Heaps lov (Herdan 1960), Menzeraths lov (Altman 1980) og First-Letter Law (Xiaoyong et al. 2017). Nærværende artikel er baseret på First-Letter Law, som er en videreudvikling af Benfords lov. Den betegnes FLL-loven eller FLL her i artiklen.

Benfords lov er et statistisk værktøj af stor almen interesse for forskere. I litteraturen finder man eksempler på mange interessante og praktiske anvendelser inden for astronomi, biologi, datavidenskab og finansiel analyse. Der knytter sig særlig stor interesse til, at loven kan bruges til at kontrollere datakvalitet. Den er især velegnet til at afsløre afvigende eller manipulerede materialer, fx AI-genererede billeder (Bonettini et al. 2020). Den videreudviklede FLL for tekst har også disse kvaliteter. Et eksempel er en undersøgelse, hvor FLL benyttes til at afsløre forekomsten af falske indlæg på Donald Trumps Twitter-konto (Guillen, Yacone & Dai 2020).

Inden for leksikografi er der tilsyneladende ingen udbredt anvendelse af Benfords lov eller FLL til undersøgelse af ordbogsindhold. Få resultater er blevet publiceret i den forbindelse. I et arbejde af Shulzinger & Bormashenko (2017) har en Benford-lignende model været brugt til almene ordbøger på 10 sprog. Så vidt vides, foreligger der på nuværende tidspunkt endnu ingen publicerede resultater, hvad angår de nordiske sprog. Derfor er det måske ikke helt forkert at hævde, at vi i Norden endnu ikke har opdaget metodens deskriptive og frem for alt analytiske potentiale.

Denne artikel forsøger at råde bod på dette ved at undersøge metodens mulige anvendelser i nordisk leksikografi, dels som en overordnet deskriptiv model af en ordbogs samlede lemmabestand, dels som en analytisk metode til at teste og sammenligne udvalgte dele af en ordbogs indhold.

2. Målsætning

Følgende hovedspørgsmål danner udgangspunkt for arbejdet:

- Er FLL i stand til at forudsige begyndelsesbogstavernes kvantitative fordeling i tre større almene skandinaviske ordbøger, og hvordan forholder de tre ordbøgers FLL-modeller sig til hinanden?
- Har FLL praktisk anvendelse ved analyse af diakrone forandringer i ordbøger?

3. Data

3.1. Materiale 1

Ordbøger	**Udgivelsesår**	**Sprog**	**Estimeret antal opslagsord**
Nudansk Ordbog	1987	Dansk	50.000
Riksmålsordboken	1977	Norsk	50.000
SAOL	1982	Svensk	115.000

Tabel 1: De tre almene ordbøger, som indgik i undersøgelsen.

3.2. Materiale 2

Ordbøger	**Sprog**	**Antal præfiksverber med *för-* og estimeret antal opslagsord i parentes**
P. O. Welander: *Svensk-dansk-norsk lommeordbog* (1846)	svensk > dansk/norsk	380 (40.000)
K. Kristensen m.fl.: *Svensk-Dansk Ordbog* (2010)	svensk > dansk	324 (50.000)

Tabel 2: De to ordbøger, der blev brugt til diakron analyse.

4. Metode

4.1. FFL-loven udtrykt i generelle matematiske termer

Den generelle formel for FLL er følgende:

$$p_i = \frac{X - (X - 1)\log_x (X - 1) - i \log_x i + (i - 1)\log_x (i - 1)}{X (X - 1) \log_x (\frac{X}{X - 1})} \quad [1]$$

p_i er FLL-værdien for et begyndelsesbogstav med hyppigheden *i*.

X er antallet af bogstaver i alfabetet eller antallet af forskellige begyndelsesbogstaver i et undersøgt delmateriale.

Log_x betegner logaritmen med basen *X*.

4.2. Praktiske beregninger af FLL-værdier i MS Excel

Hvis man benytter MS Excel, kan FLL-værdierne beregnes for et alfabet med 28 bogstaver som følger:

For *i* = 1 (dvs. det begyndelsesbogstav, som forekommer hyppigst i materialet) beregnes FLL-værdien med:

p1 = (1+27*LOG(28/27;28))/(28*27*LOG(28/27;28)) [2]

For *i* = 2 (dvs. det næsthyppigste begyndelsesbogstav i materialet) beregnes FLL-værdien med denne formel:

p_2 = (28-27*LOG(27;28)-2*LOG(2;28)+1*LOG(1;28))/(28*27*LOG(28/27;28)) [3]

Sidstnævnte formel [3] kan derefter benyttes til alle de øvrige begyndelsesbogstaver med faldende hyppighed. For fx at få FLL-værdien for det sjette-hyppigste begyndelsesbogstav udskifter man de to 2-taller med 6 og de to 1-taller med 5.

Formel [2] og [3] kan kopieres direkte ind i Excel. For at få en tilpasning til frekvenserne i det aktuelle materiale multipliceres FLL-værdierne med det totale antal ord.

4.3. Tilpasninger af FLL-beregningerne til de aktuelle materialer

FLL-værdierne blev beregnet for 29 bogstaver for den danske og norske ordbog, fordi de indeholdt ord med begyndelsesbogstavet *W*, og for 28 begyndelsesbogstaver for den svenske ordbog, der ikke havde opslagsord, som begyndte med *W*.

FLL-værdierne for præfiksverberne med *för-* blev bestemt ud fra det første bogstav i andetleddet og blev beregnet for 23 bogstaver i materiale 2. Præfikser, der begynder med *för-* (fx *förut-*, *före-*, *förbi-*), er ikke medtaget i materialet.

4.4. Optælling og estimering af antal opslagsord

De tre ordbøgers lemmabestand blev estimeret ved at optælle antallet af linjer per begyndelsesbogstav i ordbøgerne, og den gennemsnitlige forekomst af tydeligt fremhævede opslagsord per linje blev beregnet ud fra stikprøver. Det totale antal opslagsord per begyndelsesbogstav blev derefter beregnet ved at multiplicere det optalte antal linjer med det gennemsnitlige antal opslagsord per linje. Ordbogens totale antal opslagsord blev beregnet ved at summere alle begyndelsesbogstavers antal.

Præfiksverberne i *Svensk-dansk-norsk lommeordbog* og *Svensk-Dansk Ordbog* blev optalt manuelt.

4.5. Præsentationen af resultaterne

Resultaterne præsenteres i todimensionelle kurvediagrammer, hvor to linjer sammenlignes. Den ene linje repræsenterer den beregnede FLL-linje (FLL-modellen), og den anden linje repræsenterer ordantallet for den pågældende ordbog. Det kan også være to FLL-modeller, som sammenlignes, fx ved sammenligning af to ordbøger. På x-aksen angives begyndelsesbogstaverne i den rækkefølge, som bestemmes af ordantallet, i retningen fra højest til lavest. Den beregnede R^2-værdi (forklaringsgraden) angives også.

4.6. Beregninger i Excel

Hvis en ordbog skal sammenlignes med dens FLL-kurve, opretter man først to kolonner: en kolonne med begyndelsesbogstaverne og en kolonne med det antal opslagsord, der har de pågældende begyndelsesbogstaver. Derefter sorteres de to kolonner efter ordantallet, fra højest til lavest. I en tredje kolonne lægger man derefter FLL-værdierne, som beregnes ifølge formel [2] og [3]. Disse FLL-værdier multipliceres nu med det samlede antal opslagsord for ordbogen. Cellerne i de tre kolonner markeres, og man tegner et diagram ved at vælge Excels funktion for todimensionelle diagrammer i menuen.

Diagrammer for sammenligning af to ordbøgers FLL-modeller kræver lidt mere håndtering af kolonner, men princippet er det samme. Man gennemfører ovenstående for begge ordbøger, men inden multiplikation af FLL-værdierne sorteres begge ordbøger alfabetisk hver for sig. Derefter markeres alle seks kolonner og sorteres efter FLL-værdierne for den ordbog, der skal udgøre udgangsordbogen (den ordbog, som den anden ordbog testes mod). Det næste trin er at multiplicere begge kolonner med FLL-værdier med det ordantal, som udgangsordbogen har. Nu kan man tegne et diagram, hvor de to ordbøgers FLL-modeller sammenlignes

med hinanden. Til dette benyttes kolonnen med begyndelsesbogstaverne for udgangsordbogen og de to kolonner med FLL-værdier.

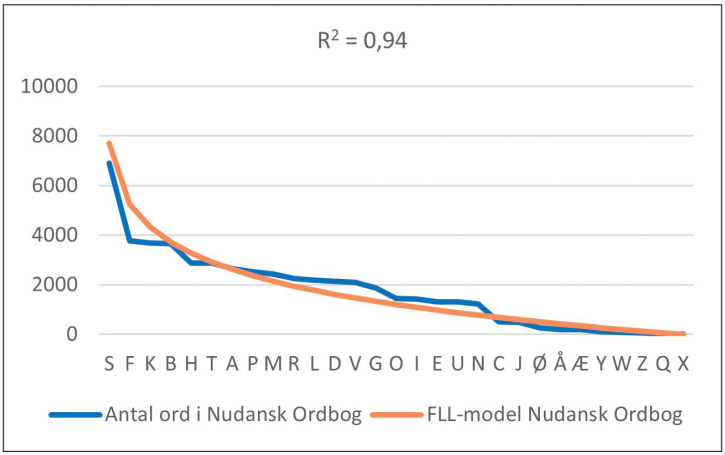
4.7. Beregninger af forklaringsgrad i Excel

Afvigelser i forhold til FLL-modellen i diagrammerne beregnes med forklaringsgraden R^2, også kaldet variationskoefficienten. I MS Excel benyttes funktionen Forklaringsgrad. R^2-værdien varierer mellem 0 og 1, hvor 1 svarer til 100 % korrelation. I nærværende artikel benyttes %-angivelsen.

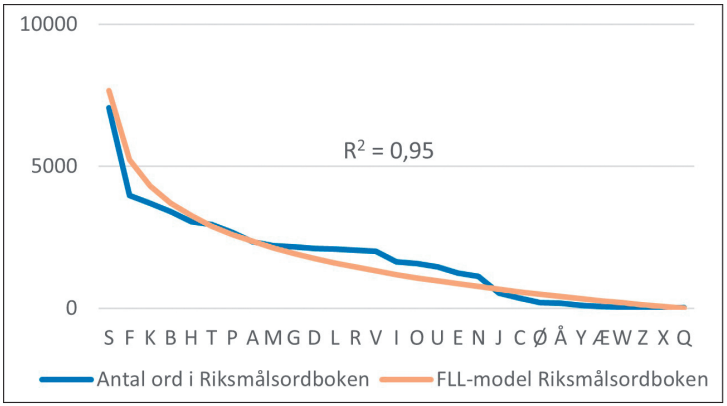
5. Resultater

5.1. *Nudansk Ordbog*, *Riksmålsordboken* og SAOL

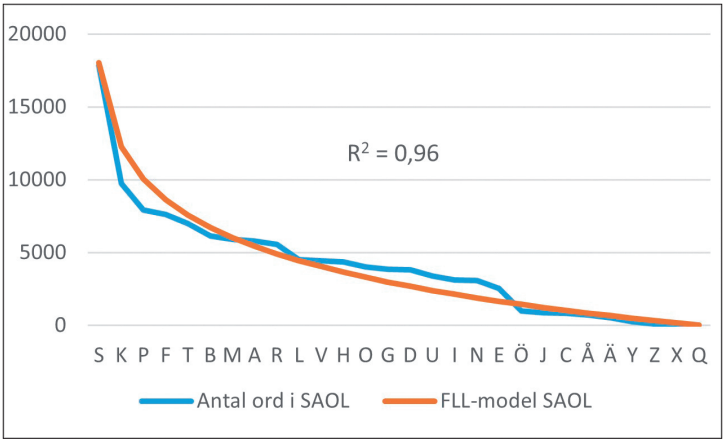
De beregnede FLL-modeller for de tre ordbøger (figur 1-3) forklarer en meget stor del af begyndelsesbogstavernes frekvenser med forklaringsgrader på 94-96 %. Der er tale om næsten identiske re-



Figur 1: *Nudansk Ordbog* og dens FLL-model.

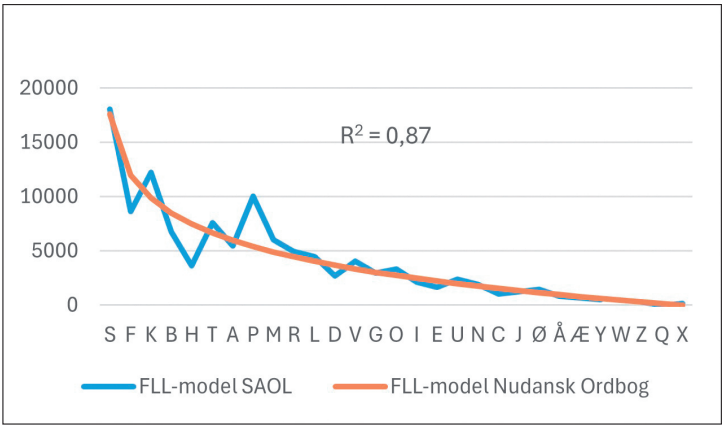


Figur 2: *Riksmålsordboken* og dens FLL-model.

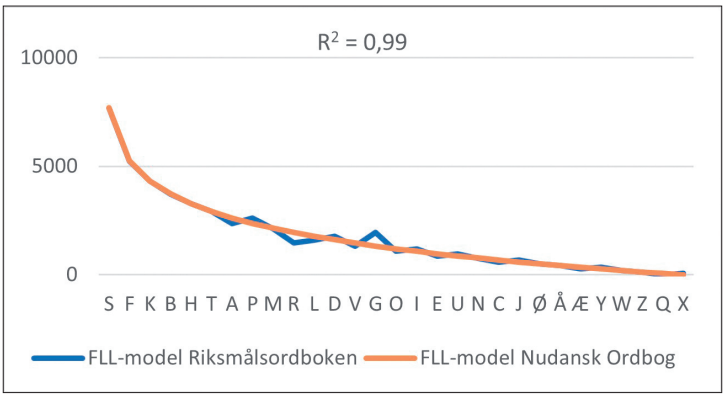


Figur 3: SAOL og dens FLL-model.

sultater, når man sammenligner diagrammerne. Selv de mindre afvigelser fra modellen mellem *S* og *H* (afvigelse under linjen) og mellem *T* og *C* (afvigelse over linjen) har det samme mønster i de tre diagrammer. Den store forskel i ordantal mellem SAOL (115.000) og de andre to ordbøger (50.000) påvirker ikke forklaringsgraden signifikant.



Figur 4: Sammenligning af FLL-modellen for SAOL med FLL-modellen for *Nudansk Ordbog*.



Figur 5: Sammenligning af FLL-modellen for *Riksmålsordboken* med FLL-modellen for *Nudansk Ordbog*.

Det betyder dog ikke, at de tre ordbøger har eksakt samme struktur. Ved sammenligning mellem *Nudansk Ordbog* og SAOL (figur 4) fremkommer der signifikante forskelle samt en lavere forklaringsgrad (87 %). Det er indikeret, at nogle af de største forskelle ses ved bogstav *H* og *P*. Umiddelbart tænker man på, at *hv-* i dansk svarer til *v-* på svensk (fx *hvid*/*vit*). Derimod er det

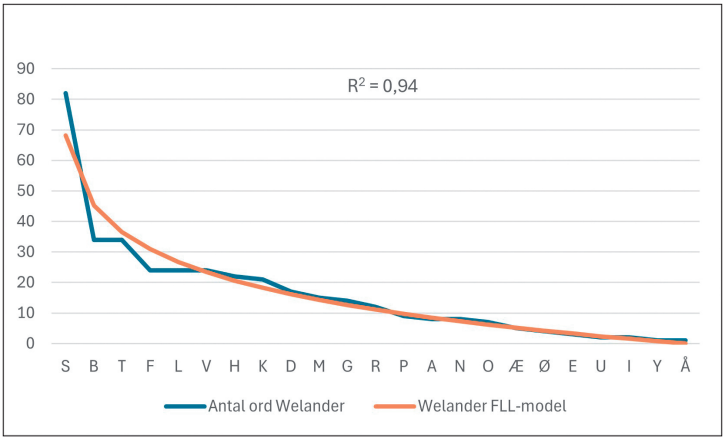
vanskeligere at forklare forskellen, når det gælder *P*. Det kan være interessant at notere, at det faktisk er muligt at sætte et tal (R^2-værdien) på den relative lighed mellem dansk og svensk ordforråd og ortografi ved at sammenligne ordbøgernes lemmafordeling.

Mellem den danske og norske ordbog (figur 5) er der betydeligt mindre forskelle end ved sammenligning mellem den danske og svenske ordbog. Forklaringsgraden er på hele 99 %.

5.2. FLL-modellen anvendt til dokumentation af diakrone forandringer

I figur 6 vises antal præfiksverber med *för-* i *Svensk-dansk-norsk lommeordbog*. Den beregnede FLL-model dækker en stor del af variationen mellem begyndelsesbogstaverne med en forklaringsgrad på hele 94 %. Det er helt på niveau med nutidens ordbøger og indikerer, at Welander har udført et genuint indsamlingsarbejde i forbindelse med ordbogen.

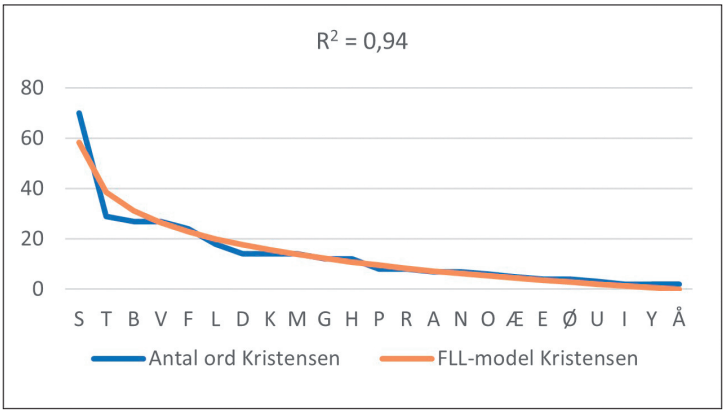
Resultatet er meget lig resultatet for den meget senere udgivne *Svensk-Dansk Ordbog* (figur 7), hvor man også ser en forklarings-



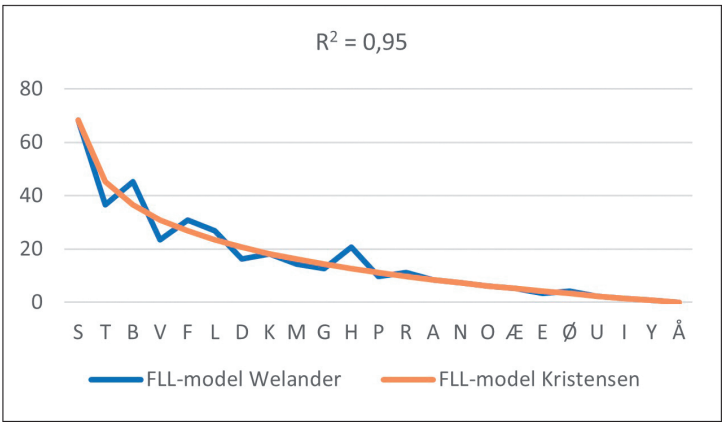
Figur 6: Antal svenske verber med præfikset *för-* i *Svensk-dansk-norsk lommeordbog* og den beregnede FLL-model.

grad på 94 %. I begge tilfælde er der kun forholdsvis små afvigelser fra FLL-modellen.

Ved sammenligningen mellem ordbøgernes FLL-modeller (figur 8) fremkommer der lidt flere og lidt mere markante afvigelser. Forklaringsgraden er dog stadig betragtelig høj, men visuelt er der større afvigelse mellem kurverne end i figur 6 og 7. Kigger



Figur 7: Antal svenske præfiksverber med *för-* i *Svensk-Dansk Ordbog* og den beregnede FLL-model.



Figur 8: Sammenligning af FLL-modellen for *Svensk-dansk-norsk lommeordbog* med FLL-modellen for *Svensk-Dansk Ordbog* med hensyn til svenske præfiksverber med *för-*.

man nærmere i de to ordbøger på de begyndelsesbogstaver, der har størst afvigelser, fx *H*, kan man se, at der her er et relativt stort indslag af forældede ord i *Svensk-dansk-norsk lommeordbog* (*förhyda*, *förhuttla*, *förhyra* og *förhäda*), der ikke er medtaget i *Svensk-Dansk Ordbog*. Denne omstændighed indikerer, at afvigelserne i diagrammet afspejler sprogforandringer i væsentlig grad.

6. Diskussion

6.1. FLL's rolle i leksikografi

Sammenfattende er det vigtigste, denne artikel viser vedrørende FLL's rolle i leksikografi, at danske, norske og svenske lemmalister har tydelige statistiske og testbare egenskaber. Det er også forbavsende ligetil at kontrollere disse egenskaber. Man behøver blot en matematisk formel og det samlede antal opslagsord. Beregninger kan dertil udføres med ringe arbejdsindsats, hvis man behersker regnearksprogrammer.

For nordiske leksikografer bør FLL være særlig interessant, fordi metoden kan give nye indfaldsvinkler i studiet af de skandinaviske nabosprogs parallelle og divergerende udvikling i ordbøger. Test af beskaffenhed og dækningsgrad kan også have betydning ved udvikling af nye ordbøger.

6.2. Der er store ligheder med andre sprog

I Shulzinger & Bormashenko (2017) har man beregnet den gennemsnitlige forklaringsgrad mellem begyndelsesbogstavers frekvens og beregnede modelværdier i ordbøger på 10 sprog (engelsk, tysk, lettisk, fransk, italiensk, spansk, hebræisk, russisk, polsk og tjekkisk) til 0,91 (91 %). Til sammenligning har de tre

her undersøgte ordbøger (dansk, norsk og svensk) et gennemsnit på 95 %, altså lidt over gennemsnittet for de 10 ovennævnte sprog. Men forskellen (4 %) er for lille til at drage nogen sikre konklusioner. I de publicerede diagrammer for tysk, russisk og fransk i samme artikel er der også en tydelig overensstemmelse med de tre skandinaviske ordbøger.

6.3. Hvordan skal man tolke ensartede afvigelsesmønstre i forhold til FLL-kurven?

Det er ikke blot de generelle kurveforløb, der har store ligheder mellem sprogene. Afvigelsesmønstrene i forhold til FLL-modellens kurvelinje har også store ligheder, ikke bare mellem dansk, norsk og svensk, men også i relation til en række andre sprog. Det indikerer, at der mangler en universal komponent i modellen på tværs af sprogene. Her er der behov for bedre forståelse for, hvad der egentlig styrer den kvantitative fordeling af begyndelsesbogstaver og årsagen til, at FLL- og Benford-modeller er så effektive til at forudsige den. I mangt og meget opleves loven som et paradoks (Fellman 2014).

6.4. Viser FLL egentlig noget nyt?

En skeptiker kan måske hævde, at FLL viser det, man allerede ved. Men hertil kan man svare, at metoden først og fremmest tilfører overskuelighed, det vil sige, at man får et overblik over den samlede lemmabestand, og at man har en model at sammenligne med. Man kan også på en praktisk måde sætte tal på det, som man normalt kun vurderer med beskrivende ord. At arbejde modelbaseret med FLL fører desuden leksikografien nærmere de videnskabsgrene, hvor samspillet mellem teori og faktiske observationer spiller en stor rolle.

6.5. Top down-tilgang med FLL

En fordel med FLL er, at den giver mulighed for at arbejde ud fra en top down-tilgang, som indebærer en struktureret og systematisk analyse af et materiale, hvor man starter med det mest generelle og gradvis zoomer ind på det mere detaljerede. Et eksempel på dette er, at man forudsætningsløst starter med at se på de største afvigelser mellem to kurver og derefter går ned på detaljeniveau og ser, hvilke lemmaer det handler om. Det kan fx ske, at forklaringen er forekomst af forældede ord eller ortografiske forskelle. Brugen af FLL-modeller bidrager til, at man umiddelbart kan identificere de mest interessante afvigelser mellem ordbøger. Samtidig er top down-princippet med til at forhindre, at man overser noget vigtigt, fx i forhold til kun at basere analysen på nogle få udvalgte typeeksempler.

6.6. FLL er ikke begrænset til det første bogstav i et ord

Da denne artikel har brugt FLL til at analysere præfiksverber i ordbøger, har den også – på en måde – testet muligheden for at bruge FLL til at undersøge sammensatte og afledte ord. Der knytter sig måske en særlig interesse til dette i Norden, fordi man i de skandinaviske sprog jo frit kan danne nye sammensatte ord i modsætning til fx engelsk. Man skal blot være opmærksom på, at der medtages lemmaer nok til, at det bliver forsvarligt at gennemføre analysen, det vil sige, at der bliver tilstrækkeligt med data til at fastsætte en rangordning af begyndelsesbogstaverne. Ved for få data kan der opstå problemer med rangordning af de lavfrekvente begyndelsesbogstaver. Antallet af lemmaer kan dog være forholdsvis lavt. I denne artikel, hvor FLL-undersøgelsen vedrørende præfiksverber var baseret på andetleddets første bogstav, blev der opnået tilfredsstillende og stabile resultater, selvom det handlede om mindre end 1 % af det samlede antal opslagsord.

6.7. Ordbøger med divergerende antal opslagsord

De tre almene ordbøger, som blev undersøgt og sammenlignet, havde ikke det samme antal opslagsord. SAOL havde to gange så mange som hver af de to andre ordbøger. FLL-metoden er dog meget robust. Det skyldes netop, at metoden er modelbaseret. Det er ordbøgernes mønstre med hensyn til begyndelsesbogstavernes indbyrdes kvantitative relationer, der analyseres. De beregninger, der foretages, er på helt sammenlignelige værdier.

6.8. Begrænsninger i metoden

En udfordring med FLL-metoden er, at den kun arbejder med antallet af ord uden hensyn til, hvilke ord det drejer sig om. Hvis man sammenligner en ældre og nyere ordbog, der har samme antal opslagsord for et begyndelsesbogstav, indeholder de måske ikke helt de samme ord. I den ene ordbog kan der være tale om indslag af forældede ord, som ikke findes i den nye ordbog. I stedet har den nye ordbog måske fået tilført nye, men helt andre ord, der antalsmæssigt svarer til de manglende forældede ord. I sådanne tilfælde vil brug af FLL-modellen derfor underestimere de reelle forskelle.

Det kan også ske, at FLL-metoden overestimerer forskelle, hvis fx et ord har *Q* som begyndelsesbogstav i en ældre ordbog, men *K* i en nyere ordbog. Man er altså nødt til at sammenligne ordbøger mere detaljeret, inden man drager mere definitive konklusioner.

Dette arbejde har i lighed med tidligere arbejder med FLL-modellen, fx Xiaoyong et al. (2017), benyttet R^2-værdien til at beskrive relationen mellem kurver. Det bør dog overvejes at finde en bedre måde at beskrive relationen mellem kurver. I figur 8 er der fx indikeret en større afvigelse visuelt mellem kurverne end i figur 6 og 7. R^2-værdien afspejler ikke dette forhold.

Ved sammenligning mellem ordbøger, hvor interessen ikke er lagt på diakrone forandringer, bør der ikke være for stort tidsrum

mellem udgivelserne. I dette arbejde sattes en grænse på højst 10 år for sammenligning mellem nordiske ordbøger.

Med hensyn til sammenligningen af de valgte nordiske ordbøger kan der gøres den reservation, at sammenligningen er problematisk, fordi SAOL er en retskrivningsordbog og de to andre nordiske ordbøger er betydningsordbøger. I hvilken udstrækning det påvirker rangordningen mellem begyndelsesbogstaver er ikke blevet klarlagt i denne undersøgelse. Derimod er forskellene i lemmaantallet sandsynligvis af mindre betydning.

Foreliggende undersøgelse har først og fremmest været en pilotundersøgelse for at se, om metoden overhovedet fungerer. Der kræves mere målrettede oplæg, hvis man skal se nærmere på, i hvilken udstrækning rangordningen af begyndelsesbogstaver kan identificere ufuldstændigheder i en ordbogs lemmabestand, og hvad der i praksis kan udledes af forskelle.

7. Konklusion

Artiklen viser, at der er særdeles god grund til at interessere sig for FLL i nordisk leksikografi. Metoden tilbyder nogle klare fordele i forhold til traditionelt leksikografisk analysearbejde.

FLL kan forudsige begyndelsesbogstavers indbyrdes relation i en ordbog, både overordnet og i sammensatte ord, men kan ikke forudsige, hvilke bogstaver det drejer sig om. Men trods denne begrænsning bør begyndelsesbogstavers rangordning betragtes som en værdifuld vidensresurse, bl.a. fordi det giver mulighed for komparative analyser. Desuden kan metoden videreudvikles, især med hensyn til korrelationsberegningen. Eftersom det handler om rangordnede serier, vil ikke-parametriske metoder formodentlig kunne finde anvendelse, fx Kendalls tau eller Spearmans rangkorrelation.

Litteratur

Ordbøger, korpusser og digitale resurser

Nudansk Ordbog (1987). Christian Becker-Christensen, Lars Heltoft, Carol Henriksen & Peter Widell (red.). 13. udgave, 3. oplag. København: Politikens Forlag.

Riksmålsordboken (1977). Tor Guttu, Kåre Skadberg & Inge Wettergreen-Jensen (red.). Oslo: Kunnskapsforlaget.

SAOL (1982) = *Svenska Akademiens ordlista över svenska språket*. 10 uppl. Stockholm: Norstedts.

Svensk-dansk-norsk lommeordbog (1846). Peter Olof Welander. Kjöbenhavn: G.H. Jægers skandinaviske Forlagshandel.

Svensk-Dansk Ordbog (2010). Kjeld Kristensen, Else Bojsen & Pernille Folkmann (red.). København: Det Danske Sprog- og Litteraturselskab & JP/Politikens Forlagshus.

Anden litteratur

Altmann, Gabriel (1980): Prolegomena to Menzerath's law. I: *Glottometrika* 2, 1-10.

Benford, Frank (1938): The law of anomalous numbers. I: *Proceedings of American Philosophical Society* 78, 551-572.

Bonettini, Nicolò, Paolo Bestagini, Simone Milani & Stefano Tubaro (2020): On the use of Benford's law to detect GAN-generated images. I: *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE. 5495-5502. <https://doi.org/10.1109/ICPR48806.2021.9412944>.

Fellman, Johan (2014): The Benford paradox. I: *Journal of Statistical and Econometric Methods* 3(4), 1-20. <https://www.scienpress.com/Upload/JSEM/Vol%203_4_1.pdf>.

Guillen, Gabriel, Christopher Yacone & Zhuoran Dai (2020): Benford's Law Applied to Twitter Data. Working paper, Harvard

University. <https://scholar.harvard.edu/files/gabrielguillen/files/final_paper.pdf> (februar 2025).

Herdan, Gustav (1960): *Type-token Mathematics*. The Hague: Mouton.

Newcomb, Simon (1881): Note on the frequency of use of the different digits in natural numbers. I: *American Journal of Mathematics* 4, 39-40.

Shulzinger, Evgeny & Edward Bormashenko (2017): On the Universal Quantitative Pattern of the Distribution of Initial Characters in General Dictionaries: The Exponential Distribution is Valid for Various Languages. I: *Journal of Quantitative Linguistics* 24(4), 273-288. <https://doi.org/10.1080/09296174.2017.1304620>.

Xiaoyong, Yan, Seong-Gyu Yang, Beom Jun Kim & Petter Minnhagen (2017): Benford's Law and First Letter of Words. I: *Physica A: Statistical Mechanics and its Applications* 512, 15. 305-315. <https://doi.org/10.1016/j.physa.2018.08.133>.

Zipf, George Kingsley (1936): *The Psycho-Biology of Language. An Introduction to Dynamic Philology*. Oxford: Routledge.

Poul Hansen
Translatør, fil dr
Box 4158
SE-227 22 Lund
poul@danska-svenska.se