

---

## OM AT STAVE ET ORD, NÅR DELE AF RETSKRIFTEN ER INDLÆRT: EN DANSK STAVEMASKINE STYRET AF SANDSYNLIGHEDER

*Preben Kihl, dr.phil.*

*Tidligere KU, DLH, DPU*

*prkihl@gmail.com*

**Abstract:** Under stavningen af ord, hvis retstavning er ukendt for den stavende, styres sammenkoblingen mellem lyde og bogstaver i komplicerede stavesystemer af sandsynligheder. Nærværende undersøgelse søger at identificere disse sandsynligheder for dansk på grundlag af stavefejlene i Noesgaards (1945) meget store datamateriale fra en københavnsk børnepopulation. En reanalyse af hans data viser, at der i typiske danske valg mellem to muligheder, f.eks. mellem bogstaverne  $\emptyset$  eller  $y$  for lyden  $[\emptyset]$  eller mellem  $+/-$  stumt  $h$  foran  $[v]$ , generelt forekommer lige mange fejl af hver type. Andre sammenhænge mellem lyde og bogstaver giver ikke anledning til stavfejl, f.eks. de initiale lukkelydes og de snævre vokalers bogstavering. På dette grundlag opstilles en dansk, stokastisk stavemaskine med sandsynlighederne  $0,5 : 0,5$  og  $1 : 1$ . Det påvises, at de fundne fejlfrekvenser næppe er korreleret med bogstavhyppighederne i danske børnetekster.

### 1. Indledning

I kybernetikken er der tradition for at kalde dynamiske systemer og deres formelle karakteristika for "maskiner", hvad enten der er tale om levende væsner, maskiner af metal eller Newtons celeste mekanik. Stavemaskinen i denne artikel er en generel model af tænkningen hos en københavnsk skoleelev mellem 9 og 15 år, der i slutningen af 1930'erne forsøgte at stave ukendte ord ved at lyddele og ved at sætte bogstaver til lydene. I komplicerede stavesystemer som det danske er der imidlertid ofte flere stavemuligheder, hvorfor bogstaveringen i ukendte ord styres af sandsynligheder i den stavendes sind. En stavemaskine med input og tilknyttede sandsynligheder kaldes også en stokastisk proces (se afsnit 3.4).

Dansk ortografi er kompliceret, fordi sproglydene ofte staves på to eller flere måder og fordi stumme bogstaver ofte, men ikke altid, forekommer i bestemte kontekster (Basbøll 2005:88-105). Dette giver stavfejl, og i ældre og nyere danske staveundersøgelser finder man derfor karakteristiske fejlstrukturer (Bleses & Thomsen 2004, Juul 2003, Kihl 1988, Noesgaard 1945). Et eksempel på en sådan

struktur udspringer af stavereglene for det korte /ɛ/-fonem med allofonerne [ɛ, æ, a], der bogstaveres som enten æ eller e på dansk.

begge	[bɛgə]	BÆKKE
jer	[jæɾ]	JÆR
melde	[mɛlə]	MÆLLE
æske	[ɛsgə]	ESKE
tænde	[tɛnə]	TENDE
træt	[tɾæt]	TRET

I ordene, der er stavet af et 8-årigt københavnsk barn (Kihl 1988), stemmer /ɛ/-fejlene (og de øvrige fejl) overens med det danske stavesystem forstået som reglerne for, hvordan lydene principielt set kan repræsenteres ortografisk. Et andet eksempel fra samme undersøgelse illustrerer stavningen af stumt d efter bogstaverne l, n, r finalt. Den generelle danske staveregel siger, at der efter lydene [l, n, ɾ] med eller uden stød kan forekomme et d.

fald	[falʔ]	FAL
mord	[moɾʔ]	MOR
pind	[penʔ]	PEN
hal	[halʔ]	HALD
fjer	[fjeɾʔ]	FJERD
hen	[hɛnʔ]	HEND

Barnets regel er klar nok: Efter [l, n, ɾ] er der mulighed for et stumt d. Af ovenstående eksempler kan man derfor slutte, at barnet kender stavereglene, men endnu ikke har indlært retstavningen af ordene.

I en standardteori som to-rutemodellen, hvis empiriske grundlag hovedsageligt er stavfejl begået af mennesker med hjerneskade, finder ordstavning sted i to forskellige systemer (Rapsack et al. 2009, Vandermosten et al. 2012). Ord, der er lært udenad og befinder sig i det ortografiske ordforråd, staves i en direkte staverute, medens ukendte ord og nonsensord staves i en indirekte fonologisk rute. Da ordene i eksemplerne ovenfor indeholder fejl og derfor må være konstrueret af den stavende, antages disse ord at være stavet i den fonologiske rute. Heri segmenteres input fra et ukendt staveord eller et nonsensord i lyde. Disse lydsegmenter kobles derpå sammen med bogstaver i en fonem-grafem komponent, hvorfra ordet går videre til outputkomponenter, der enten staver mundtligt eller skriftligt. Reglerne for sammenkobling – stavereglene i eksemplerne ovenfor – antages at befinde sig i fonem-grafem komponenten. Imidlertid har staveundersøgelser i komplekse systemer ikke givet valide, empiriske tal for, hvor ofte lyde fejlstaves på de forskellige måder, som et givet stavesystem giver mulighed for. Ej heller for, hvor

ofte stumme bogstaver fejlagtigt udelades eller indsættes i relevante kontekster. Sådanne fejltal kan fortolkes som stavesandsynligheder i den stavendes hjerne, dvs. i to-rutemodellens fonem-grafem komponent, men er også interessante for den neurale netværksmodel for stavning (Houghton & Zorzi 2010).

Den engelske ortografi er kompliceret, og Baxter & Warringtons undersøgelse (1987) af en engelsk hjerneskadet dysgrafiker, der primært stavede via den fonologiske rute, har tal af relevans for problemstillingen. Men tallenes validitet begrænses af, at de stammer fra en caseundersøgelse. De to forskere sammenlignede endvidere patientens stavning med lyd-skrift forekomsthypighederne i en engelsk fonetisk ordbog i et forsøg på at forstå, hvorfra de korrekte og fejlstavede bogstaver stammede. Sammenligningen viste, at der generelt var "little evidence to support the hypothesis that KT [patienten] used the most frequent graphemic correspondence at the expense of other graphemic correspondences".

I Kihl (1988) findes der optællinger af eksempelvis hvor mange Æ-fejl, der begås i forhold til E-fejl for /ε/-fonemet eller hvor mange stumme D'er, der fejlagtigt udelades i forhold til D'er, der fejlagtigt indsættes. Imidlertid er undersøgelsen ligesom Baxter & Warringtons et casestudie.

I Noesgaards staveundersøgelse (1945) findes fejltal, der belyser problemstillingen. Noesgaard beskriver og optæller imidlertid ikke sine fejltal på grundlag af forholdet mellem bagvedliggende lyde og ukorrekte bogstaver, men alene ud fra forholdet mellem korrekte og ukorrekte bogstaver. Fejlkategoriene kaldes henholdsvis "bogstavforvekslinger", f.eks. æ>e og omvendt, og "udeladte og indsatte bogstaver", f.eks. l>ld og omvendt.

Disse tre arbejder har været inspirationskilder til nærværende undersøgelse, fordi de har beskæftiget sig med at finde tal, der kan belyse de sandsynligheder, som mennesker benytter, når de i komplicerede stavesystemer staver ukendte ord via den fonologiske rute. Endvidere dækker undersøgelserne sandsynligvis langt de fleste fejlmuligheder i de pågældende sprog. Andre undersøgelser i komplekse stavesystemer kunne utvivlsomt have været inddraget. På grund af problemerne i de refererede undersøgelser besluttede jeg at genoptælle dele af Noesgaards undersøgelse på et moderne grundlag. Undersøgelsen er enestående på grund af sin størrelse, idet Noesgaard og hans hjælpere i to undersøgelser i slutningen af 1930'erne foranstaltede, at 2 x ca. 10.000 københavnske skoleelever hver stavede 100 ord, i alt ca. 2 millioner ord. I de to undersøgelser indgik elever fra henholdsvis 3. til 5. klasse og fra 6. til 8. klasse, herefter benævnt retskrivningsundersøgelse RI og RII. Efter dataindsamlingen foretog Noesgaard datareduktion, idet lister med rigtighedsprocenter over 98 % og under 60 % blev sorteret fra, sådan at hvert ord i undersøgelsens materiale på i alt 3918 ord efterfølgende var blevet stavet 300 gange, i alt ca. 1,2 millioner ord. Ord materialet stammede fra en række stavelister for 2. til 8. skoleår og var ikke kontrolleret mht. variabler som hyppighed, længde, grammatisk kategori osv., der nu anses for at være relevante for sværhedsgraden.

Den ny undersøgelse har således til formål at bestemme stavesandsynlighederne i dansk på grundlag af udvalgte data fra Noesgaard og at opstille en kognitiv model af en dansk staver – en stavemaskine.

## 2. Metode

Min nye optælling skete på grundlag af Noesgaards datalister (1945: 68-116), idet de to undersøgelser og aldersgrupper, RI og RII, blev holdt adskilt i optællingerne. Listerne blev gennemgået ord for ord med henblik på at finde fejltallene for en vis type af stavfejl, f.eks. hvor mange fejl det tryksvage finale [ɐ] medførte i RI. Dette gav to talrækker, der viste, hvor mange gange [ɐ] var blevet fejlstavet som henholdsvis RE og RER og i hvilke ord, f.eks. ”kreaturer” KREATURE, ”desværre” DESVÆRRER. Disse fejlfordelinger blev analyseret statistisk, i det omfang de rapporterede data tillod en analyse. Data er ikke normalfordelt, hvorfor en binomialtest er den eneste mulige statistiske signifikansberegning af eventuelle forskelle mellem fordelingerne. Imidlertid er en khi-i-anden test kun mulig under en række urealistiske forudsætninger: At alle ord har samme fejlsandsynlighed for et givet barn, at alle børn staver et givet ord med samme fejlsandsynlighed, og at der er stokastisk uafhængighed mellem et barns stavfejl, når det f.eks. begår fejl af typen HÆST og GEST. Da ingen af disse forudsætninger er opfyldt i data, er khi-i-anden statistik ikke anvendelig.

Tabel 1. En oversigt over genoptalte fejltyper med eksempler og hypotetiske bagvedliggende lyde i Noesgaards undersøgelse

Udtalte lyde	
Kortvokalerne [ɛ, æ, a],[ø, œ, œ], [e]	Fejlstavet E/Æ, Ø/Y, E/I, f.eks. ”bedst” BÆDST, ”jævn” JEVN ”blødt” BLYDT, ”lyst” LØST ”skik” SKEK, ”elever” ILEVER
Lukkelydene [b, d, g] medialt, i dækket udlyd og udlyd	Fejlstavet P/B, T/D, K/G, f.eks. ”dybt” DYPT, ”prop” PROB ”færdig” FÆRTIG, ”alt” ALD ”lugt” LUKT, ”dukke” DUGGE
Tryksvagt finalt [ɐ]	Fejlstavet RE/RER/(ER)*, f.eks. ”tårer” TÅRE, ”føre” FØRER, men ikke ”uhyre” UHYER
[ɑ̃]-diftongens to komponenter	fejlstavet A/I/E og J/G f.eks. ”plejer” PLAJER, ”nej” NIJ, ”majs” MEJS ”tegne” TEJNE, ”fejl” FEGL
Stumme bogstaver	
D efter finalt [l], [n], [ɳ] +/- stød	fejlstavet +/-D, f.eks. ”al” ALD, ”fuld” FUL ”den” DEND, ”hånd” HÅN ”ler” LERD, ”bord” BOR
H før initialt [j], [v]	fejlstavet +/-H, f.eks. ”just” HJUST, ”hjerne” JERTE ”vis” HVIS, ”hvis” VIS

\* Fejlen ER for [ɐ] forekommer meget sjældent i data og kun i konteksten [-ɐ], f.eks. "vandre" [vandɐ] VANDER, sandsynligvis når en udtale uden [ɐ] ligger bag.

Det var en væsentlig præmis for genoptællingen af stavfejlene, at optællingen skulle baseres på nogle hypotetiske bagvedliggende lyde i ordmaterialet på grundlag af min egen udtale. Et eksempel vil belyse forskellen mellem Noesgaards og min optællingsmetode: I ord som "vil", "gængæld" og "egne" karakteriserede Noesgaard fejlene i VEL, GÆNGELD og EJNE som henholdsvis i>e, e>æ/æ>e og g>j med tilhørende fejltal.

I den nye optælling beskrives de anførte fejl som henholdsvis [e]>E, [ɛ] >Æ/E og [ɪ]>J med tilhørende fejltal.

Konsonanter og stumme bogstaver var uproblematisk i optællingen, medens Grønnum (2009:59-61) blev anvendt ved udredningen af vokalkvaliteterne og deres retstavning.

Hvordan kan der argumenteres for, at jeg anvender min egen udtale som udgangspunkt for at rekonstruere elevernes og lærernes udtale af ordmaterialet i Noesgaards undersøgelse? Lærernes udtale spiller også en rolle, idet de dikterede staveordene for eleverne.

Jeg er født ca. to år efter afslutningen på Noesgaards dataindsamling, og voksede op i københavnsområdet indtil 15 års alderen. Ifølge Brink & Lund (1974:20-21) er mit sprog næsten rent højkøbenhavnsk rigsmål og svarer derfor til sproget hos en del af børnepopulationen. Men det er en metodisk svaghed, at jeg kun har teoretisk kendskab til lavkøbenhavnsk, idet elevpopulationen og lærerne har talt begge de to københavnske dialekter. Dertil kommer jo tilflytterbørn fra Lolland og Fyn m.m.

Spørgsmålet er imidlertid, om problemet er reelt. Medmindre man har lydoptagelser og transskriptioner, er det stavfejl, der giver information om lydene bag et skrevet ord. Hvis to bagvedliggende dialekter forårsagede forskellige fejltyper, burde det kunne identificeres i Noesgaards resultater og i den ny optælling. Noesgaard skriver intet om sådanne forskelle, og i mine optællinger er der kun ét eksempel af en vis størrelse, hvori hypotetiske bagvedliggende udtaler, der afviger fra min egen udtale, efter alt at dømme optræder i Noesgaards data. Se Juul (2012) i denne forbindelse. Eksemplet drejer sig om vokalen i ord som "ret", som jeg udtaler [ɐd], men som dele af Noesgaards elevpopulation må have udtalt [ɐd] ligesom i nudansk. Dette medførte, at ord som "omtrent" og "græs" blev fejlstavet som OMTRÆNT/OMTRANT, GRES/GRAS. I hele datasættet forekom E- og Æ-fejl lige ofte, medens A-fejl forekom halvt så ofte. I den statistiske analyse af /ɛ/, se næste afsnit, er A-fejlene holdt ude pga. antagelsen om, at den bagvedliggende lyd ikke er en allofon af /ɛ/, men netop [ɑ]. A-fejlene er imidlertid også data. I afsnit 3.4 om "stavemaskinen" vises det vha. eksemplet "træt", hvordan fejlstavning på grundlag af to forskellige udtaler af et ord kan analyseres.

Et hovedformål med optællingen var som nævnt at finde fejlfordelingerne for stavningen af lyde og stumme bogstaver og at analysere disse. Noesgaards egne

analyser angiver det kvantitative forhold mellem udeladte, indsatte og forvekslede bogstaver og fejlgennemsnit, f.eks. at bogstavet ø fejlagtigt blev erstattet med y 507 gange i 46 ord med et gennemsnit på 11 fejl per ord i RI. Desuden har han oversigter over svære ord med mere end 50 fejl, f.eks. at ordet "trøste" havde 78 y-fejl i RI. Fordelingstypen, hvor mange y-fejl de øvrige ord gav anledning til, og hvilke [ø]-fonemer og -allofoner, der indgår i ordene, rapporteres ikke af Noesgaard.

Forskellene i optællingsmetode og optællingskriterier – i denne undersøgelse optælles som nævnt kun udvalgte dele af Noesgaards materiale – medfører, at taloversigterne i de to undersøgelser bliver forskellige og ikke umiddelbart sammenlignelige.

Her optælles alene kortvokalfejl i stavelser med fuldtryk samt fejl ved tryksvagt [ɐ] finalt. Noesgaard, derimod, inkluderer også lang- og stødvokalfejl ved lydene/bogstaverne æ, e, ø, slår e/i fejl ved [ɑɪ]-diftongen sammen med de øvrige vokalfejl og optæller [ɐ] overalt. Begrundelsen for at udelukke fejl ved lang-, stød- og potentielle stødvokaler er, at der sædvanligvis kun er en fejlmulighed ved disse vokaler og at deres inklusion i optællingen ville skævvride fordelinger med to valgmuligheder for stavningen af en given lyd.

Eksempler på fejl, der inkluderes af Noesgaard, men ikke i den nye optælling, er æ/e-fejl i "bære" BERE, "Legeme" LÆGEME, "værre" VERE, "Sjæl" SJEL, "lænede" LENEDE, "Næb" NEB, "hæve" HEVE, "Eventyr" ÆVENTYR, y/ø-fejl i "følte" FYLTE, "brølede" BRYLEDE, "lød" LYD, "R/røg" RYG, Frøken FRYKEN samt e/i-fejl i "enige" INIGE, "bed" BID, "dejlig" DIJLIG, "sig" SEG, "direktør" DIRIKTØR osv. Her optælles kun tryksvagt [ɐ] finalt, medens Noesgaard optæller fejl ved tryksvagt [ɐ] overalt samt ved [ɶɐ] og [ɸɐ], f.eks. "Varerne" VARENE, "yderligere" YDERLIGER, "Digtere" DIGTER og "vandrør" VANDER, "Forældre" FORÆLDER.

Fejltallene ved [ɑɪ]-diftongen er i de to undersøgelser kun delvis sammenlignelige for 1. komponentens A/E-fejl pga. fejl som "ret" RAT, og ikke sammenlignelige for I/E-fejlene, der som nævnt slås sammen med I/E-fejl ved [e]-vokalen. Den manglende sammenlignelighed gælder også J/G-fejl ved [ɑɪ]-diftongens 2. komponent, der hos Noesgaard er sammentalt med de øvrige diftongers fejl, f.eks. "møjsom" MØGSOM, "bevægelse" BEVÆJELSE osv., medens eksempelvis fejl som "Leje" LEGGE er inkluderet her, men ikke i Noesgaards optælling.

Hvor Noesgaard optæller fejl med stumt d og stumt h overalt, f. eks. "værste" VÆRDSTE, "falder" FALER, "Bunden" BUNEN, "hændte" HÆNTE, "hente" HENDTE, "ihjel" IJEL, "enhver" ENVER, "Erhverv" ERVERV, "alvor", ALHVOR, optælles disse fejltyper blot finalt og initialt i denne undersøgelse.

Hvad angår de uaspirerede lukkelyde [b, d, g], dvs. b/p-, t/d- og k/g-fejl, er der i denne undersøgelse i forhold til Noesgaards set bort fra fejlene i ord som "Bordplade" BORDBLADE, "værd" VÆRT, "voldsomt" VOLTSOMT, "stødte" STØTTE, "stationen" STADIONEN, "tag" TAK, "strengt" STRENKT, "seksten" SEGSTEN, "langt", LANKT, "Punkt" PUNGT osv. Endvidere er de sjældne lukkelydsfejl i forlyd og dækket forlyd udeladt i den ny optælling, f.eks. "Prop" BROP,

”Spanden” SBANDEN. Det samme gælder t/d-fejl i ord som ”løbet” pga. vekslen mellem [d/ð] i udtalen. Inkluderet i den ny optælling i forhold til Noesgaards er fejl som ”Kapital” KABBITAL, ”Blikket” BLIGET, ”mødtes” MØDDES, osv.

Ordene i eksemplerne ovenfor fra Noesgaards lister er her kun rekonstruerede mht. den relevante fejltype, da Noesgaard ikke selv giver eksempler på fejlstavede ord, kun bogstav-bogstav regler. Eksemplerne kan derfor indeholde flere forskellige andre fejl, der ikke er medtaget her.

### 3. Resultater

#### 3.1. Analysen af fejlfordelingerne

##### 3.1.1 Fejlstavede lyde

Af hensyn til overskueligheden er tallene fra RI og RII i dette afsnit adderet i tabellerne. Når der forekommer udviklingsmæssige forskelle mellem RI og RII, vil det blive omtalt i kommentaren til tabellerne.

Tabel 2. Oversigt over fejlstavninger af kortvokalerne [ɛ, æ, a], [ø, œ, œ,], [e] og [ɐ] i RI + RII

Fejltype	Ord med fejl	Antal stavfejl	Fejl per ord	Standardafvigelse	Sandsynlighed
/ɛ/>Æ	313	4443	14,19	16,69	0,47
/ɛ/>E	342	5508	16,11	21,75	0,53
/ø/>Ø	101	1784	17,66	16,89	0,63
/ø/>Y	74	773	10,45	13,34	0,37
[e]>E	194	2178	11,23	13,42	0,60
[e]>I	32	236	7,38	7,13	0,40
[ɐ]>RE	44	1500	34,09	29,27	0,63
[ɐ]>RER	81	1891	23,35	19,5	0,37

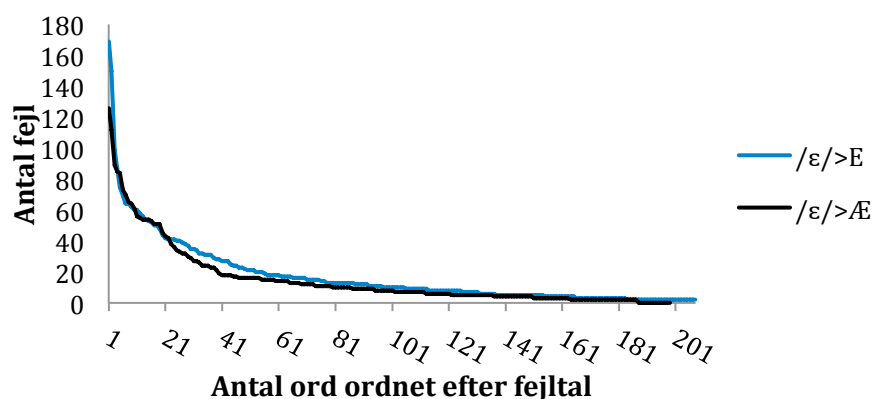
Forekomsten af forskellige antal ord med fejl indenfor en fejltype i tabel 2 og i de øvrige tabeller skyldes, at Noesgaards data ikke er styret eksperimentelt mht. antallet af ord, hvori bestemte lyde indgår. Sandsynligheden for, at en lyd fejlstaves på den ene, anden (eller tredje) måde og i senere tabeller for, at et stumt bogstav fejlagtigt indsættes eller udelades, fremgår af yderste, højre søjle.

Fejlhyppighederne ligger til grund for beregningen af sandsynlighederne. Da der er forskel på antallet af ord i samples, beregnes sandsynlighederne i tabel 2 og senere tabeller ud fra gennemsnittet af fejl per ord, f.eks. er sandsynligheden for en forkert Æ-stavemåde for lyden /ɛ/ 0,47 i tabel 2, idet  $14,19 / (14,19 + 16,11) = 0,47$ .

Tabel 2 viser, at fejlstavninger med bogstavet E har en lidt højere sandsynlighed end fejlstavninger med bogstavet Æ for /ɛ/ fonemet, men at sandsynlighederne er næsten fifty-fifty. Forholdet mellem E og I fejl for lyden [e] er 2 : 3. Forholdet mellem Ø og Y fejl for /ø/ er 2 : 1 i RI, men ca. 1 : 1 i RII. Forholdet mellem fejlene RE og RER for tryksvagt [ɐ] er ca. 1 : 2. Den upræcise sprogbrug ved

karakteristikken af de kvantitative forskelle og ligheder skyldes som nævnt, at signifikansberegninger ikke er mulige. De to største fejlfordelinger i Noesgaards data vedrører stavningen af /ε/. Fejlene forekom i 655 ord, der hver især blev stavet 300 gange. Udgangsmaterialet udgør således i alt 196.500 stavede ord, hvori stavefejl af typen ”hest”> HÆST, ”gæst”>GEST forekom.

Figur 1. Kurver over fejlstavningen af kort /ε/ i RI



Stavefejl i ord af typen ”gæst” > GEST er ordnet i den stiplede kurve med faldende fejltal fra venstre mod højre. Den fuldt optrukne kurve, der viser stavefejlene i ord som ”hest”> HÆST, er ordnet på samme vis.

Det ses, at de to kurver har samme forløb på trods af den manglende styring af ordmaterialet i undersøgelsen. Figur 1 (og figur 2 senere i dette afsnit) skal grafisk illustrere, at der er tale om ensdannede, lige store fordelinger. Se også tabel 6-8 i afsnit 3.1.3 samt argumentationen i starten af afsnit 3.4.

Tabel 3. Oversigt over fejlstavninger af lukkelydene [b, d, g] medialt, i dækket udlyd og udlyd i RI + RII

Fejltype	Ord med fejl	Antal stavefejl	Fejl per ord	Standardafvigelse	Sandsynlighed
[b]>B	68	984	15,14	20,8	0,47
[b]>P	50	855	17,1	20,45	0,53
[g]>G	183	2684	14,67	21,52	0,44
[g]>K	126	2379	18,84	21,65	0,56
[d]>D	137	799	5,83	6,51	0,29
[d]>T	54	784	14,52	19,27	0,71

I RI er der ca. dobbelt så mange B og G fejl som P og K fejl, medens fejlfordelingen for stavningen af lydene [b, d, g] i RII er ca. 0,5 : 0,5. Overvægten af fejlstavninger med T for [d] i det samlede resultat skyldes især høje fejltal ved ordene ”bold” og ”fodbold” i RI måske pga. homofoni med ”bolt”.



Tabel 4. Oversigt over fejlstavninger af [ɑ̃]-diftongens 1. og 2. komponent i RI + RII

Fejltype	Ord med fejl	Antal stavfejl	Fejl per ord	Standard-afvigelse	Sandsynlighed
[ɑ̃]>A	35	204	6,18	7,14	0,15
[ɑ̃]>I	6	40	6,67	6,5	0,16
[ɑ̃]>E	6	215	28,5	35,59	0,69
[ɪ̃]>J	43	1190	27,67	32,34	0,50
[ɪ̃]>G	61	1692	27,74	24,81	0,50

Ved fejlstavningen af [ɑ̃]-diftongens 2. komponent [ɪ̃] anvendes mulighederne J og G lige ofte i RI og RII.

Da alle fejlmulighederne A, E, I anvendes ved stavningen af [ɑ̃], kan et ukendt ord fejlstaves på to måder. Dette er der to eksempler på i data, nemlig ”dejlig” og ”nej”, fejlstavet DAJLIG/DIJLIG og NAJ/NIJ. De tre sandsynligheder for fejlstavning af [ɑ̃] er jævnt fordelt med værdier omkring 0,33 i RI. I RII derimod forekommer bogstavet E som fejl 3 gange så hyppigt som A + I pga. ordene ”majjs” og ”kaptajn”, der tilsammen fejlstaves med E 157 gange.

### 3.1.2 Fejlstavede stumme bogstaver

Tabel 5. Oversigt over fejlstavninger af stumt d efter finalt [l, n, ɳ] samt af stumt h foran initialt [j, v] i RI + RII

Fejltype	Ord med fejl	Antal stavfejl	Fejl per ord	Standard-afvigelse	Sandsynlighed
[l]>L	21	270	12,86	9,30	0,48
[l]>LD	32	448	14	10,79	0,52
[n]>N	41	524	12,78	15,45	0,45
[n]>ND	30	466	15,53	16,72	0,55
[ɳ]>R	11	571	51,91	46,36	0,82
[ɳ]>RD	31	343	11,06	13,62	0,18
[j]>J	12	142	11,83	12,21	0,54
[j]>HJ	16	163	10,19	9,40	0,46
[v]>V	43	1653	38,44	33,48	0,66
[v]>HV	113	2200	19,47	31,57	0,34

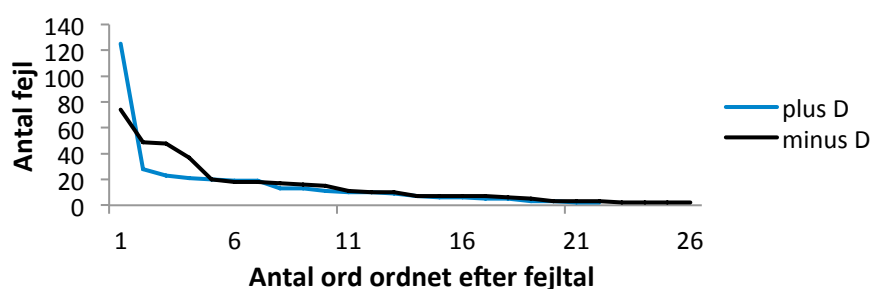
Det fremgår af tabel 5, at ukendte ord med finalt [l, n] i omtrent lige stort omfang enten tillægges eller mangler et d, medens chancen for et udeladt d efter [ɳ] er ca. 4 gange højere end for et forkert indsat d. En mulig forklaring på forskellen kunne være, at lærerne, der dikterede testordene til Noesgaards børnepopulation, har haft varierende udtaler af f.eks. ”værd” som enten [væɳʔ] eller [væʔɳ]. Hvordan dette har

påvirket elevernes stavning er uklart, men et gæt kunne være, at når udtalen har været [væ<sup>ʔ</sup>ɐ], kunne dette i højere grad have medført et udeladt D end ved udtalen [væɸ<sup>ʔ</sup>].

Fejlsandsynlighederne for stumt h før initialt [j] i RI og RII og for [v] i RII er jævnt fordelt, men ikke før initialt [v] i RI, hvor sandsynligheden for et udeladt h er 0,71.

Nedenfor vises fordelingerne af stavfejl i RII ved stumt d efter [l, n, ɸ] grafisk.

Figur 2. Kurver over fejlstavningen af stumt d efter finalt [l, n, ɸ] i RII



Kurverne i figur 2 har et ens forløb, selvom datamaterialet på  $49 \times 300 = 14.700$  stavede ord er langt mindre end det, der lå bag kurverne over stavfejl ved /ɛ/ i figur 1.

### 3.1.3 Overblik og generalisering

Noesgaards datareduktion, se afsnit 2, medførte utvivlsomt, at de fleste stavemæssige vildskud blev sorteret fra. De resterende stavfejl i data kaldes i stavelitteraturen fejl, der er fonologisk plausible i en ortografi, fejl, der lyder som mål-ordet, når de læses op, lydbevarende fejl, overgeneraliseringer osv. Kihl et al. (2000) gennemgår forskellige forskeres varierende betegnelser for de stavfejl, der er systematiske og velordnede i forhold til en given ortografi. Dvs. fejl, der er i overensstemmelse med et givet stavesystems "fornuftige" muligheder for fejlstavning i modsætning til fejl, der er uordnede og fonologisk uplausible (Elbro 2015: 85-86).

Det umiddelbare indtryk af tallene i de foregående tabeller er, at stavfejl af den ene og anden type indenfor en fejlkategori i ret høj grad har lige stor sandsynlighed for forekomst på trods af den manglende eksperimentelle kontrol i Noesgaards undersøgelse. Måske de store tal udjævner forskelle? Kan en sådan intuition bestyrkes på trods af de forskelle i sandsynligheder, der også er fundet? Det virker logisk, at stavfejlene ved de undersøgte stumme bogstaver h og d kan slås sammen som fejl, hvori der enten udelades et stumt bogstav eller indsættes et stumt bogstav i en relevant kontekst. En sædvanlig fortolkning af udeladelsesfejlene går på, at det stumme bogstav ikke skrives, fordi der ikke udtales en lyd. Derfor kaldes denne fejltypen her "fonetisk", fordi der er overensstemmelse mellem (ingen) lyd og (intet)

bogstav. Sådanne udeladelsesfejl kan imidlertid fortolkes på to måder: Som lydstavning med kun en stavemulighed uden noget alternativ eller som resultatet af, at den stavende er klar over at stavesystemet rummer to muligheder i en given kontekst, enten et stumt bogstav eller ej, og tilfældigvis vælger den forkerte mulighed. I modsætning hertil er forkerte indsættelser af relevante stumme bogstaver på korrekte steder i ord sikre tegn på, at der er sket en indlæring af stavesystemets muligheder hos den stavende. Denne fejltipe kaldes her for ”ikke-fonetisk”.

Følgelig er alle fejl med de stumme bogstaver h og d i Noesgaards undersøgelse kategoriseret som fonetiske og ikke-fonetiske og adderet i tabel 6, hvori tallene fra både RI, RII og RI + RII er medtaget.

Tabel 6. Oversigt over udeladte og indsatte stumme bogstaver i RI, RII og RI + RII

Aldersgruppe og fejltipe	Ord med fejl	Antal stavefejl	Fejl per ord	Sandsynlighed
RI				
Fonetiske	86	2408	28	0,64
Ikke-fonetiske	142	2282	16,07	0,36
RII				
Fonetiske	42	752	17,90	0,52
Ikke-fonetiske	80	1338	16,73	0,48
RI + RII				
Fonetiske	128	3160	24,69	0,6
Ikke-fonetiske	222	3620	16,31	0,4

Tabel 6 repræsenterer 6780 stavefejl i 350 ord i RI og RII. Da hvert ord er stavet 300 gange, er det bagvedliggende datamateriale i alt 105.000 stavede ord. Som det ses, er der en overvægt af udeladte stumme bogstaver i RI og i det samlede resultat, medens der er en mere jævn fejlfordeling i RII.

I Tabel 7 er stavefejlene ved de udtalte lyde kategoriseret og adderet på følgende grundlag: Når et barn staver ”pynte” som PØNTE, må man antage, at barnet perciperer, at der er en [ø]-lyd i ordet og derfor anvender ø-bogstavet, hvis bogstavnavn eller -lyd stemmer overens med [ø]. Sådanne fejl kunne man klassificere som ”fonetiske fejl”, fordi der er overensstemmelse mellem lyd og bogstavlyd/bogstavnavn. Når det samme barn et øjeblik senere staver ”løftede” som LYFTEDE, er det klart, at barnet har lært, at [ø] i ord også kan skrives med et y-bogstav i dansk. Fejltypen kunne man derfor i analogi til kategoriseringen af fejl ved stumme bogstaver kalde ”ikke-fonetisk”. En forkert anvendelse af et y-bogstav ved stavningen af en [ø]-lyd er ligesom indsættelsen af et relevant stumt bogstav i en korrekt kontekst et sikkert kendetegn på indlæring hos den stavende. Derimod kan fonetiske fejl ved udtalte lyde ligesom fonetiske fejl ved stumme bogstaver fortolkes som enten lydstavning eller som resultatet af indlæring baseret på stavesystemets egenskaber.

Tabel 7. Oversigt over fejlstavninger af udtalte lyde i RI, RII og RI + RII

Aldersgruppe og fejltype	Ord med fejl	Antal stavfejl	Fejl per ord	Sandsynlighed
RI				
Fonetiske	619	8829	14,26	0,45
Ikke-fonetiske	448	7844	17,51	0,55
RII				
Fonetiske	455	5437	11,95	0,45
Ikke-fonetiske	303	4485	14,80	0,55
RI + RII				
Fonetiske	1074	14.266	13,28	0,45
Ikke-fonetiske	751	12.329	16,42	0,55

Tabel 7 repræsenterer 26.595 stavfejl i 1825 ord i RI og RII. Da hvert ord er stavet 300 gange, er det bagvedliggende datamateriale 547.500 stavede ord. Det ses, at sandsynlighederne for fonetiske og ikke-fonetiske fejl ligger omkring 0,5 : 0,5 med en vis overvægt af ikke-fonetiske fejl i både RI og RII.

I tabellen indgår ikke fejlstavningerne af [e], da det ikke har været muligt at kategorisere disse fejl meningsfuldt som henholdsvis fonetiske og ikke-fonetiske. A-fejl ved [ɑ̃]-diftongens 1. komponent er klassificeret som fonetiske, E/I fejl som ikke-fonetiske.

I tabel 8 er stavfejlene ved stumme bogstaver og udtalte lyde slået sammen i overensstemmelse med kategoriseringerne i tabel 6 og 7.

Tabel 8. Oversigt over fejlstavninger af stumme bogstaver og udtalte lyde i RI, RII og RI + RII

Aldersgruppe og fejltype	Ord med fejl	Antal stavfejl	Fejl per ord	Sandsynlighed
RI				
Fonetiske	705	11.237	15,94	0,48
Ikke-fonetiske	590	10.126	17,16	0,52
RII				
Fonetiske	497	6189	12,45	0,45
Ikke-fonetiske	383	5823	15,20	0,55
RI + RII				
Fonetiske	1202	17.426	14,50	0,47
Ikke-fonetiske	973	15.949	16,31	0,53

Tabel 8 repræsenterer 33.375 stavfejl i 2175 ord i RI og RII. Da hvert ord er stavet 300 gange, er det bagvedliggende datamateriale 652.500 stavede ord. Stavesandsynlighederne i tabel 6-8 vil blive anvendt ved konstruktionen af en dansk stavemaskine i afsnit 3.4.

Den relativt ens fordeling af fonetiske og ikke-fonetiske fejl i tabellerne kunne skyldes, at forskellige børn anvender forskellige stavestrategier i en stor populationsundersøgelse i et kompliceret stavesystem. Nogle elever udelader måske ret konsekvent stumme bogstaver og anvender bogstavlydene ved stavningen af sproglyde i deres stavfejl, medens andre elever måske i stort omfang gør det modsatte og begår ikke-fonetiske fejl. Dette kan ikke umiddelbart afvises. Men da stavfejlene i caseundersøgelser som især Kihl (1988:149-152), men også Baxter & Warrington (1987), Beauvois & Dérousné (1981) og Kihl et al. (2000), generelt set fordeler sig på samme vis som fejlene i tabel 8, er der næppe empirisk belæg for antagelsen.

### **3.2. Stammer de fundne sandsynligheder fra bogstavhyppighederne i børnetekster?**

Analysen af fejlfordelingerne fra Noesgaards data har vist, at eleverne i typiske danske valgsituationer med to muligheder for bogstavering stavede relativt tilfældigt, dvs. ca. 0,5 : 0,5. En oplagt forklaring på disse tilfældige valg mellem alternative muligheder kunne være, at populationen var blevet påvirket af, at bogstavhyppighederne i danske børnetekster også er 0,5 : 0,5. At de fundne fejlfordelinger afspejler forekomsthypighederne af bogstaver i input til Noesgaards elever.

Til belysning af denne problemstilling blev Maegaard & Ruus' (1981) hyppighedsliste over ord i danske børnetekster anvendt som sammenligningsgrundlag. Heri foretog jeg optællinger i de første og hyppigste 1000 ord, der dækker 81,25 % af de faktiske ordforekomster i Maegaard & Ruus' tekster. En optælling af endnu 1000 ord ville blot have givet en yderligere dækning på ca. 5 % og blev derfor ikke foretaget. I tabel 9 bliver forholdene mellem stavfejl fra tabellerne 2-5 sammenlignet med de samme forhold mellem bogstavhyppigheder hos Maegaard & Ruus.

Tabel 9. Forhold mellem stavfejlhyppigheder hos Noesgaard i RI + RII og bogstavhyppigheder hos Maegaard &amp; Ruus (1981)

Lyde og stumme bogstaver	Forhold mellem stavfejl	Bogstaver og bogstavhyppigheder	Forhold mellem bogstaver
/ɛ/>E/Æ	1,08:1	e:æ 26.312:3445	7,64:1
[e]>E/I	1,52:1	e:i 10.386:15052	0,69:1
/ø/>Ø/Y	1,69:1	ø:y 1753:457	3,84:1
[ɐ]>RE/RER	1,50:1	re:rer 2600:89	29,21:1
[ɑ]>E/I/A	4,61:1,08:1	e:a:i 5257:2876:40	1,83:0,01:1
[ɪ]>J/G	1:1	j:g 1153:6991	0,16:1
[g]>K/G	1,28:1	k:g 8722:1847	4,72:1
[d]>T/D	2,49:1	t:d 19.311:1504	12,84:1
[b]>P/B	1,13:1	p:b 2471:1826	1,36:1
[l,n,ŋ]-D/+D	1,38:1	-d:+d 11.608:3070	3,78:1
-H/+ H[v]	1,97:1	-h:+h 10.271:2946	3,49:1
-H/+H[j]	1,16:1	-h:+h 5011:619	8,10:1

På overfladen er der i tabel 9 væsentligt større forskel og variation i forholdene mellem bogstaver end i forholdene mellem stavfejl. En beregning af korrelationskoefficienten ud fra de forholdstal i hver række, der ikke er 1, f.eks. tallene 1,08 og 7,64 i første række (ved lyden [ɑ] er tallene 4,61 og 1,83 anvendt), viser da også en negativ, ikke signifikant korrelation tæt ved 0 ( $r = -0,06$ ,  $p < 0,2$  tohalet).

Korrelationen på grundlag af tallene i tabel 9 godtgør, at der næppe er nogen sammenhæng mellem stavfejlhyppighederne i de udvalgte data fra Noesgaard og bogstavhyppighederne i danske børnetekster. Dette resultat svarer til Baxter & Warringtons (1987), nemlig at deres patients fejllhyppigheder ikke var korreleret med de engelske bogstavhyppigheder. Overensstemmelsen foreligger på trods af forskellige sammenligningsgrundlag: Dels bogstavhyppighederne i danske børnetekster baseret på antallet af faktisk forekommende ord eller "tokens", dels bogstavhyppighederne i en engelsk fonetisk ordbog med ordtyper eller "types".

Det kan imidlertid ikke være rigtigt at sammenligne bogstavhyppigheder fra en ordbog med stavfejlhyppigheder. Altså eksempelvis at bruge det danske digitale system "bogstavlyd.ku.dk", der i princippet svarer til Baxter & Warringtons fonetiske ordbog, i en sammenligning. Både børn og voksne læser jo tekster, men kun i begrænset omfang ordbøger. Endvidere er bogstavhyppighederne meget forskellige i ordbøger og tekster. I ordbøger forekommer ordene kun én gang, hvorimod de 100 hyppigste ord i Maegaard & Ruus' børnetekster forekommer mellem 9371 og 341 gange og dækker 55,82 % af alle bogstavforekomster i børneteksterne. Når børn lærer at læse og stave vha. læsebøger og småtekster, kigger de på mange eksempler på de samme børneord, ikke på ordtyper i ordbøger.

I komplicerede ortografier udleder børn på ukendt vis hurtigt lydene bagved bogstaverne og udleder de steder, hvor stumme bogstaver kan forekomme. Dette

indikerer danske og internationale undersøgelser af stavning (Dick & Cohn 2013, Juul 2012, Kihl 1988, 1993, Noesgaard 1945, Viise et al. 2011).

Desuden argumenterer Treiman & Kessler (2014:66-68) for betydningen af implicit statistisk indlæring, idet "children take advantage of [the world's] predictability by learning how often events occur ... and co-occur ... they develop a kind of mental statistics" og refererer en række undersøgelser til støtte herfor.

Men dette afsnits korrelationsanalyse har vist, at Noesgaards elevers stavning ikke blev styret af de statistiske inputforhold i danske børnetekster, idet børnenes stavfejlhyppigheder sandsynligvis ikke afspejler bogstavhyppighederne i deres læsestof.

### 3.3. Noesgaards elever havde imidlertid indlært dele af retskriften i ordene

Som vist forekom Noesgaards elevers stavfejl ved de danske en-flere relationer mellem lyd og bogstav. Derimod forekommer stavfejl yderst sjældent ved lyde, hvor stavesystemets relationer er en-en (se også Kihl 1988:143-144).

De initiale lukkelyde [b, d, g, p, t, k], inklusive [b, d, g] i initiale s-konsonantgrupper, staves praktisk taget altid korrekt som b, d, g, p, t, k. Men den korrekte stavning af disse lukkelyde er tydeligvis indlært, fordi lydene [b, d, g] medialt, i dækket udlyd og udlyd meget hyppigt fejlstaves. Noesgaards population gør forskel på den initiale stavning af de uaspirerede lukkelyde overfor stavningen medialt, i dækket udlyd og udlyd. I RII staves s-konsonantgrupper stort set korrekt overalt.

Konsonanterne [m, n], [f, h], [v, j, ʋ], [l] samt halvokalerne [ð] efter fuldvokal og [ɸ] staves ligeledes korrekt af Noesgaards elever. Finalt [əð] staves derimod hyppigt forkert som D. Populationen kender forskellen på stavningen af [ð] efter fuldvokaler og schwa. Der er nogle ekstra problemer knyttet til konsonanterne, idet man som staver skal lære, at konsonanter i indlyd skal repræsenteres med to bogstaver efter kortvokal, men med et konsonantbogstav efter langvokal. Endvidere skal det indlæres, at der kan være knyttet stumme konsonantbogstaver til lydene [v, j], [l, n, ɸ] og [d, s] enten før eller efter lyden. Sådanne valgsituationer medfører mange stavfejl. Men isoleret set staves disse konsonant- og halvokallyde praktisk taget korrekt. Konsonanterne [ɲ, ʃ, s], der hver især har mange stavemåder, tages op i diskussionsafsnittet.

Vokalsystemet er rimelig enkelt at karakterisere. Da lydene [i, y, u, o, ə] samt /a/ med få undtagelser både i kort, lang og stødt form er entydigt repræsenteret i det danske system, staves de stort set korrekt. De øvrige lang- og stødvokaler volder af samme grund ikke de store problemer med undtagelse af langvokalerne [æ:, ɔ:], der har to stavemåder, nemlig e/æ og å/o. Resten af vokalsystemet inklusive halvokalerne [ɥ, ɹ] giver fejl.

Det er væsentligt, at der i Noesgaards ordlister ikke forekommer fremmedord, franske låneord og navne.

Det er således muligt at opstille en tentativ oversigt over, hvad der staves rigtigt og forkert af velfungerende danske elever fra 3. til 8. klasse på grundlag af Noesgaards (og Kihls, 1988) undersøgelser.

### 3.4. En dansk, stokastisk stavemaskine baseret på stavfejl og korrekt stavning i en normalpopulation på ca. 20.000 københavnske børn fra 3. til 8. klasse omkring 1940

Den samlede oversigt i tabel 8 gav de gennemsnitlige sandsynligheder 0,53 : 0,47 med en overvægt af ikke-fonetiske fejl. Oversigten over de stumme bogstavers stavfejl i tabel 6 gav sandsynlighederne 0,6 : 0,4 med en overvægt af fonetiske fejl især i RI, medens tabel 7, der dækker de udtalte lyde, gav sandsynlighederne 0,55 : 0,45 til fordel for de ikke-fonetiske fejl i både RI og RII. Denne forskel mellem stumme bogstaver og udtalte lyde er en nyhed, hvis den er valid. Og i en stavemaskine kunne man uden videre anvende de faktiske resultater fra tabellerne 6 og 7.

I denne undersøgelses optællinger indgår der imidlertid lidt over 5 gange så mange ord og knap 4 gange så mange stavfejl ved de udtalte lyde som ved de stumme bogstaver. Man må derfor antage, at inklusionen i optællingen af flere stumme bogstaver i flere positioner, især af stumt d foran s og t, af stumt d medialt efter l, n, r, af stumt g efter snævre vokaler, og af ikke-initialt stumt h foran j og v, ville medføre, at de gennemsnitlige sandsynligheder i Noesgaards materiale ville nærme sig 0,5 : 0,5. Disse værdier vil derfor blive anvendt efterfølgende.

Med stavesandsynlighederne på plads kan stavemaskinen introduceres. Maskinmetaforen stammer som nævnt fra kybernetikken (Ashby 1968:161-163, Novikov 2016).

Modelbygningen i dette afsnit har til formål at præsentere den kognitive mekanisme, som Noesgaards population anvendte, når de stavede et ord, hvis retstavning de ikke kendte. Denne mekanisme eller maskine udgør fonem-grafem komponenten i to-rutemodellen for stavning, der blev omtalt i indledningen.

En ”stokastisk overgangsmatrice” er en ofte anvendt kybernetisk modellering af en statistisk styret proces. I matricerne nedenfor står 0 for usandsynlige stavemåder, sandsynligheden 1 stammer fra afsnit 3.3, medens sandsynlighederne 0,5 : 0,5 næppe behøver yderligere introduktion. De nøjagtige værdier fra RI og RII og RI + RII kan naturligvis indsættes i modellerne. Dette vil der blive givet et eksempel på.

I tabel 10 forudsættes det, at stoveoperationen ”find et passende bogstav til en given lyd eller udelad/indsæt et passende stumt bogstav i en given kontekst” bruges på lydrækken fra venstre mod højre efter endt lydsegmentering (Holligan & Johnston 2005).

Tabel 10. Stokastisk overgangsmatrice for stavningen af ”hvem”

lyde/ bogstaver	[v]	[ɛ]	[mʔ]
v	0,5	0	0
hv	0,5	0	0
æ	0	0,5	0
e	0	0,5	0
m	0	0	1



Matricen for ”hvem” angiver, at lyden [v] enten fejlstaves v eller hv med sandsynligheden 0,5. Men [v] har sandsynligheden 0 for at blive skrevet som enten æ, e eller m. Og så fremdeles. Lyden [mʔ] transformeres til bogstavet m med sandsynligheden 1. Tabel 10 viser, at stavemåderne VEM, HVEM, VÆM, HVÆM er lige sandsynlige under forudsætning af, at staveren ikke kender ordets retstavning.

I tabel 11 anvendes de faktisk fundne sandsynligheder fra RI på ordet ”hvem”.

Tabel 11. Stokastisk overgangsmatrice for stavningen af ”hvem” i RI

lyde/ bogstaver	[v]	[ε]	[mʔ]
v	0,71	0	0
hv	0,29	0	0
æ	0	0,48	0
e	0	0,52	0
m	0	0	1

I RII ligger fejlsandsynlighederne for initialt [v] og [ε] tæt på 0,5.

Det næste eksempel belyser bl.a. stavningen af en initial s-konsonantgruppe, der staves som en enhed med sandsynligheden 1 for at blive stavet korrekt, se afsnit 3.3.

Tabel 12. Stokastisk overgangsmatrice for stavningen af ”skole”

lyde/ bogstaver	[sg]	[o:]	[l]	[ə]
sk	1	0	0	0
o	0	1	0	0
l	0	0	1*	0
e	0	0	0	1

\*Bogstavsfordobling forekommer

Sandsynligheden for at stave ordet ”skole” korrekt er 1, hvad enten den stavende har lært ordets retstavning eller ej.

For at vise hvordan varierende udtaler af et staveord kan inkorporeres i analysen, er ordet ”træt” analyseret i matricen i tabel 13.

Tabel 13. Stokastisk overgangsmatrice for stavningen af ”træt”

lyde/ bogstaver	[t]	[ʁ]	[a]	[ɑ]	[d]
t	1	0	0	0	0
r	0	1	0	0	0
æ	0	0	0,5	0	0
e	0	0	0,5	0	0
a	0	0	0	1	0
t	0	0	0	0	0,5
d	0	0	0	0	0,5

Som forklaret i metodeafsnittet udtales vokalen i ordet ”træt” utvivlsomt på to måder i Noesgaards population. Når den hypotetiske bagvedliggende udtale er [a], gælder stavesandsynlighederne i 4. søjle. Når den er [ɑ], gælder sandsynligheden i 5. søjle som antaget i afsnit 3.3, idet forudsætningen som altid er, at ordenes retstavning ikke er kendt.

Der er i to-rutemodellen ingen forskel på stavningen af rigtige ord, hvis stavemåde er ukendt for den stavende, og på stavningen af nonsensord. Nonsensord segmenteres og tilordnes bogstaver i fonem-grafem komponenten på den fonologiske rute på samme måde som ukendte ord. Til illustration af stavemaskinens hypotetiske funktionsmåde analyseres stavningen af et nonsensord i tabel 14.

Tabel 14. Stokastisk overgangsmatrice for stavningen af nonsens-ordet [vønʔ]

lyde/ bogstaver	[v]	[ø]	[nʔ]
v	0,5	0	0
hv	0,5	0	0
ø	0	0,5	0
y	0	0,5	0
n	0	0	0,5
nd	0	0	0,5

Matricen skal bl.a. demonstrere det danske stavesystems mangfoldighed ved stavningen af en simpel 3-lydsstruktur på det indlæringsstadium, hvor Noesgaards elever befandt sig. Det teoretiske antal af mulige stavemåder for lydene skal ganges med hinanden, følgelig er der  $2 \times 2 \times 2 = 8$  mulige og alle lige sandsynlige stavemåder, f.eks. HVYND eller VØN.

Afslutningsvis en analyse af et lidt længere ord fra Noesgaards dataliste i RI (1945:94). I hans egen notation: Værksted..., h+v8, æ>e40,..., k>g8, e>æ3,... Opstillingen angiver som nævnt forholdet mellem rigtige og forkerte bogstaver og viser først, at h fejlagtigt indsættes foran v 8 gange. Resten af opstillingen angiver forkerte erstatninger, dvs. at æ fejlstaves som e 40 gange, at k fejlstaves som g 8

gange og at e fejlstaves som æ 3 gange. Prikker angiver fejltal mht. versaler og ”andre fejl” m.m.

Table 15. Stokastisk overgangsmatrix for ordet ”værksted” [væꞥgsdēð]

lyde/ bogstaver	[v]	[æ]	[ꞥ]	[g]	[sd]	[ε]	[ð]
v	0,5	0	0	0	0	0	0
hv	0,5	0	0	0	0	0	0
e	0	0,5	0	0	0	0	0
æ	0	0,5	0	0	0	0	0
r	0	0	1	0	0	0	0
g	0	0	0	0,5	0	0	0
k	0	0	0	0,5	0	0	0
st	0	0	0	0	1*	0	0
e	0	0	0	0	0	0,5	0
æ	0	0	0	0	0	0,5	0
d	0	0	0	0	0	0	1

\*Gælder for RII, i RI er konsonantgruppen [sd] kun optalt initialt.

#### 4. Konklusion og diskussion

Stavesandsynlighederne fra Noesgaards data viser, at der for hans københavnske elever kun var statistisk forskel på stavningen af de rigtige og forkerte bogstaver i et ukendt ord. Populationens stavning af ukendte ord blev styret af nogle enkle sandsynligheder, der næppe hang sammen med input, dvs. bogstavhyppighederne i de tekster, som eleverne læste. Sådanne sandsynligheder er ikke tidligere blevet bestemt i komplekse stavesystemer. Staveindlæringen i de to-tre første skoleår medførte efterfølgende, at en-flere relationer mellem lyde og bogstaver og mellem stumme bogstaver og deres kontekster gav fejl, medens en-en relationer mellem lyde og bogstaver gav korrekt stavning. Noesgaards elever havde indlært dele af retskriften i ordene. I dette scenarium bliver stavefejlene alene markører, der for iagttageren viser, at der foregår noget interessant.

Disse oplysninger er meget anvendelige i standardmodellen for stavning – to-rutemodellen – hvori lydene i ukendte staveord efter segmentering tilordnes bogstaver fra venstre mod højre i en fonem-grafem komponent i overensstemmelse med sandsynligheder som dem, der er fundet denne undersøgelse. Fonem-grafem komponenten repræsenteres i denne artikel vha. stokastiske overgangsmatricer af staveord, såkaldte ”kanoniske repræsentationer” (Ashby 1968). Disse stokastiske overgangsmatricer er modeller af det, der kaldes ”assembled spelling” hos Martin & Barry (2012), der efterlyser sådanne modeller.

Som nævnt i indledningen kaldes de ord, der ikke går gennem stave-maskinens stokastiske apparat – den stavendes udenad indlærte forråd af staveord – for ”det ortografiske ordforråd” i to-rutemodellen (Rapscak et al. 2009, Tainturier et

al. 2006). Det antages i litteraturen, at gode stavere næsten udelukkende anvender disse ord, når de staver. Ordnes kognitive natur er ukendt, men en sædvanlig udlægning af de empiriske fund er, at de i løbet af læsestaveudviklingen bliver indlært som helheder af en art, f.eks. som ”staveordbilleder” (se f. eks. Ehri 2009, Martinet et al. 2004, Tainturier, Schiementz & Leck 2006).

Når eleverne i Noesgaards undersøgelse i valgsituationer stavede med sandsynligheder omkring 0,5 : 0,5, må det logisk følge, at ca. halvdelen af børnenes gæt gav stavefejl, halvdelen korrekt stavning. Dette åbner muligheden for en foreløbig bestemmelse af størrelsen af et ortografisk ordforråd. Når et barn eksempelvis har en fejlprocent på 20 i en stavetest, kan man multiplicere de 20 % med 2 = 40 % ud fra en antagelse om, at barnet har gættet rigtigt i halvdelen af de valgsituationerne, det er blevet stillet overfor. Dvs. at 40 % af ordene i testen er blevet konstrueret og stavet selvstændigt af barnet – ordene er vandret ad den fonologiske rute igennem stavemaskinen. Resten af ordene, 60 %, udgør følgelig det ortografiske ordforråd. Men da danske ord ofte indeholder adskillige fejlmuligheder (Noesgaard 1945:117-118), beror en nøjagtig måling af størrelsen af et ortografisk ordforråd på, hvilke ord, der indgår i et givet testmateriale. Når fejlmulighederne i forskellige staveord er henholdsvis 0, 1, 2, 3, 4 osv., er sandsynlighederne for korrekt stavning henholdsvis 1, 0,5, 0,25, 0,125, 0,0625 osv. Det er derfor nødvendigt at bestemme fejlmulighederne i hvert enkelt ord i en given test. Dette muliggør en beregning af, hvor mange af de rigtigt stavede ord, der er stavet korrekt ved tilfældighedernes spil, hvorefter en endelig beregning af størrelsen af det ortografiske ordforråd kan finde sted.

Hvad angår lyde med flere end to stave- og fejlmuligheder er undersøgelsens resultater relativt sparsomme. Ganske vist var der en jævn fordeling af stavefejl i RI (men ikke i RII) ved stavningen af lyden [a] i [aɪ]-diftongen med tre fejlmuligheder, nemlig E/A/I, men data er få.

Analysen af stavefejl ved tryksvagt finalt [ɐ] viste endvidere, at kun to af tre potentielle fejlmuligheder blev udnyttet i større omfang, nemlig RE og RER, idet fejlen ER kun forekom i forbindelse med udtalen [-ɐ], sandsynligvis når en udtale uden [ɐ] lå bag, f.eks. ”vandre” [vændɐ] VANDER.

Som omtalt indgår der ikke fremmedord, franske låneord og navne i Noesgaards ordlister. Lyden [s] har, udover et eventuelt associeret stumt d, adskillige stavemåder i dansk, hvoraf kun bogstaverne s og c forekommer i ordlisterne og som stavefejl. Bogstavet c står kun initialt i Noesgaards data, hvorfor der kun er optalt stavefejl ved initialt s. I RI + RII fejlstaves [s] 17 gange som S i 3 ord med initialt c, f.eks. SYKEL, og 83 gange som C i 7 ord, f.eks. CEPTEMBER. Heraf kan man se, at nogle af Noesgaards elever har en skjult viden om, at [s] kan staves på to måder og at andre bogstaver for [s] næppe eksisterer i børnenes bevidsthed, når staveordet er ukendt.

Lydene [ŋ, ɕ] har flere eksotiske stavemåder. En kursorisk gennemgang viser, at lydene staves som ng og sj og stort set korrekt pga. ordmaterialet, undtagelse ”station”. Konsonantgruppen [-ŋg(-)] med stavemåden nk er ikke undersøgt.

Gælder resultaterne i denne undersøgelse blot for stavningen i en københavnsk elevpopulation i slutningen af 1930'erne? Dette vil blive undersøgt ved en sammenligning med Kihl (1988) og med danske staveundersøgelser efter 2000.

Noesgaards undersøgelse og Kihls caseundersøgelse af et 7-8 årigt barn, der talte højkøbenhavnsk, er begge survey undersøgelser, der dækker hovedparten af mulige fejltypen hos københavnske børn på tre forskellige udviklingstrin. En systematisk sammenligning mellem de to undersøgelser fejltypen, fordelingen af fejl indenfor fejltypen samt de relative fejlprocenter, dvs. en given fejltypes procentuelle andel af alle fejlene i et datasæt, viste kun mindre forskelle. Disse forskelle må tilskrives aldersforskellen, idet Kihls forsøgsperson var 1-2 år yngre end Noesgaards yngste elever.

De offentliggjorte undersøgelser om dansk stavning efter år 2000 arbejder med udvalgte problemstillinger. Nogle af disse studier vil her blive diskuteret i forhold til nærværende undersøgelse mht. metodik, klassifikationssystemer og resultater.

Juul (2003, 2005) argumenterer mod den arbitrære opdeling i "vokal- og konsonantforvekslinger" og foreslår et system med tre forskellige grader af forudsigelighed for de fonologisk plausible fejl. Det mest forudsigelige niveau, "enkeltlydsniveauet", svarer til denne undersøgelses en-en relationer mellem lyd og bogstav. De to følgende niveauer, "lydfølgeniveauet" eksemplificeret ved kvist-kvest og "ordniveauet" eksemplificeret ved blæst-blest, svarer til nærværende undersøgelses en-flere relationer. I en undersøgelse af 104 københavnske elever fra 4.-6. klasse fandt Juul, at børnene stavede lyden [e] med kontekstuelle begrænsninger mere korrekt end lyden [ɛ] med ordspecifik stavning. Dette kunne ligne en modstrid i forhold til denne undersøgelses ens fordeling af stavfejl ved de samme vokaler, men er det ikke, da Juul beskæftiger sig med stavefærdighed, der måles i fejlprocenter. Ifølge Juul medfører kontekstforhold i dansk, at ord med [e] staves mere korrekt end ord med [ɛ], men ifølge nærværende undersøgelse ord vælger den stavende tilfældigt mellem bogstaver, når man alene ser på antallet af stavfejl ved [e] og [ɛ] i fejlstavede ord. Dette indikerer, at [e]'s kontekst faciliterer input til det ortografiske ordforråd, dvs. staveindlæringen, men ikke påvirker stavningen af ikke indlærte staveord.

En beregning på grundlag denne undersøgelses tal i tabel 2 for RI + RII støtter Juuls fund på et uafhængigt grundlag, idet fejlprocenten og fejl per fejlstavet ord for lyden [e] er 3,56 og 9,31, medens de tilsvarende tal for /ɛ/ er 5,06 og 15,15.

En anden pointe er, at intet peger på, at sproglige forskelle mellem Noesgaards og Juuls elever skulle spille en rolle.

I 2004 publicerede Bleses & Thomsen en undersøgelse af samspillet mellem talt og skrevet sprog i en fynsk børnegruppe på 20 elever i 4. klasse. Der var fire fejkategorier, hvoraf de to, nemlig "lydrelaterede fejl" og "andre fejl", svarer til denne undersøgelses fonetiske og ikke-fonetiske fejl. Bleses & Thomsen sammenlignede imidlertid transskriptioner af elevernes udtale med deres stavning og var på dette grundlag i stand til at relatere fynske former og schwa reduktioner til stavningen, f. eks. "stirrede" STIERD og "kiggede" KIKKED. Se også Juul (2012) i

denne forbindelse. Ejendommeligt nok forekommer finale stavfejl, der kan fortolkes som schwa reduktion, praktisk taget ikke i Noesgaards vældige materiale. Schwa reduktion var nemlig ifølge Brink & Lund (1974:26) udbredt i københavnsk omkring 1940 og [ə] forekom sjældent undtaget ”snævert bundet til distinkt tale”. Lærerne, der underviste Noesgaards elever og dikterede testordene, har utvivlsomt talt pædagogisk tydeligt til børnene.

Juul & Sigurdsson (2005) sammenlignede stavningen af initiale og mediale konsonanter og konsonantgrupper hos elever i 3. og 4. klasse fra Reykjavik og København vha. nonsensord. Et resultat var, at staveindlæring i den komplekse danske ortografi sammenlignet med den mere transparente islandske ortografi var en art handicap, idet entydige lyd-skrift sammenhænge i konsonantgrupper, herunder s-lukkelydsgrupper, og initiale konsonanter blev stavet signifikant mere forkert af de danske elever end af de islandske. Som nævnt staves initiale lukkelyde og s-lukkelydsgrupper i Noesgaards undersøgelse praktisk taget altid korrekt i RI og RII (fejlprocent for initiale s-lukkelydsgrupper i RI 0,25). Noesgaards elevers stavning af s-konsonantgrupperne kunne derfor ligne stavningen af ortografiske helheder, dvs. at to eller flere bogstaver i kontekst altid staves korrekt. Men pga. Juul & Sigurdssons fund kan man indse, at Noesgaards elever tidligere i deres staveudvikling må have opløst og fejlstavet de initiale s-lukkelydsgrupper. Under forudsætning af, at stavningen af ukendte ord og nonsensord som antaget i tabel 14 og i to-rutemodellen styres af sandsynlighederne fundet i denne undersøgelse, kan man endvidere slutte, at københavnske elever omkring 1940 var dygtigere til at stave end Juul & Sigurdssons københavnske elever omkring 2005.

Data i nærværende undersøgelse er indsamlet i en københavnsk børnepopulation omkring 1940, men i lyset af sammenligningen med de moderne danske staveundersøgelser foreslås det, at artiklens stokastiske stavemaskine også er en generalisation af nutidige danskeres indlærte stavekompetence mellem 3. og 8. klasse i det danske skolesystem og måske resten af livet (Kihl et al. 2000).

Bemærk dog, at maskinen ikke dækker ordblind stavning (Peterson & Pennington 2012) eller stavningen hos ganske unge, såkaldte børnestavere uden de store skriftsprogserfaringer (Hagtvet 1989, Quellette & Senechal 2008).

Det virker også rimeligt, at undersøgelsens resultater kunne gælde for stavningen i andre komplekse alfabetiske sprog.

## Litteratur

- Ashby, W. R. (1968). *An introduction to cybernetics*, Methuen.
- Basbøll, H. (2005). *The phonology of Danish*, Oxford University Press.
- Baxter, D. M. & E. K Warrington (1987). Transcoding sound to spelling: Single or multiple sound unit correspondences? *Cortex*, vol. 23: 11-18.
- Beauvois, M.-F. & J. Dérusné (1981). Lexical or orthographic dysgraphia, *Brain*, vol. 104: 21-50.

- Bleses, D. & P. Thomsen (2004). The acquisition of spoken forms and written words: An empirical study of opacity in the speech/reading/writing interface in Danish, *Written Language & Literacy*, vol. 7: 79-99.
- Brink, L. & J. Lund (1974) *Udtaleforskelle i Danmark. Aldersbestemte – geografiske – sociale*, Gjellerups Forlag.
- Brink L. & J. Lund (1975). *Dansk Rigsmål I. Lydudviklingen siden 1840 med særlig henblik på sociolekterne i København*, Gyldendal.
- Dick, N. & A. C. Cohn (2013). A review of spelling acquisition: Spelling development as a source of evidence for the psychological reality of the phoneme, *Lingua*, vol. 133: 213-229.
- Ehri, L. C. (2009). Learning to read words: Theory, Findings, and Issues, *Scientific studies of Reading*, vol. 9: 167-188.
- Elbro, C. (2015). *Læsevanskeligheder*, Gyldendal.
- Grønnum, N. (2009). *Fonetik og fonologi. Almen og dansk*, Akademisk forlag.
- Hagtvet, B. E. (1989). *Skriftsprogsudvikling gennem leg*, Gyldendal.
- Holligan, C. & R. S. Johnston (2005). Spelling errors and phonemic segmentation ability: the nature of the relationship, *Journal of Research in Reading*, vol. 14: 21-32.
- Houghton, G. & M. Zorzi (2010). Normal and impaired spelling in a connectionist dual-route architecture, *Cognitive neuropsychology*, vol. 20: 115-126.
- Juul, H. (2003). Typer af stavefejl og typer af stavefærdighed, *Danske Talesprog*, vol. 4: 59-74.
- Juul, H. (2005). Knowledge of context sensitive spelling as a component of spelling competence: Evidence from Danish, *Applied psycholinguistics*, vol. 26: 249-265.
- Juul, H. (2012). Utdaledistinkthed og stavefærdighed hos danske børn, *Danske Talesprog*, vol. 12: 96-109.
- Juul, H. & B. Sigurdsson (2005). Ortography as a handicap? A direct comparison of spelling acquisition in Danish and Icelandic, *Scandinavian Journal of Psychology*, vol. 46: 263-272.
- Kihl, P. (1988). *En skoledrengs staveregler. En undersøgelse af strukturen i og motiverne til 700 stavefejl i 1200 ord stavet af et 7-8 årigt københavnsk barn i løbet af 2. klasse*, Odense Universitetsforlag.
- Kihl, P. (1993). Visible phonemes – or how to extract an inventory of phonemes from a corpus of spelling errors, *IRAL: International review of applied linguistics in language teaching*, vol. 31: 189-204.
- Kihl, P., K. Gregersen & N. Sterum (2000). Hans Christian Andersen's spelling and syntax: Allegations of specific dyslexia are unfounded, *Journal of learning disabilities*, vol. 33: 506-519.
- Leask, A. & F. Hincliffe (2007). The effect of phonological awareness intervention on non-word spelling ability in school-aged children: An analysis of qualitative change, *Advances in Speech-Language Pathology*, vol. 9: 226-241.
- Maegaard, B. & H. Ruus (1981). *Hyppige ord i danske børnebøger*, Gyldendal.
- Martin, D. H. & C. Barry (2012). Writing nonsense: the interaction between lexical

- and sublexical knowledge in the priming of nonword spelling, *Psychonomic bulletin and review*, vol. 19: 691-698.
- Martinet, C., S. Valdois & M. Fayol (2004). Lexical orthographic knowledge develops from the beginning of literacy acquisition, *Cognition*, vol. 91: B11-B22.
- Noesgaard, A. (1945). *Fejltyper i dansk Retskrivning*, Fr. Bagges kgl. Hofbogtrykkeri.
- Novikov, D. A. (2016). *Cybernetics: From past to future*. Studies in systems, decision and control. Springer Link.
- Peterson, R. L. & B. F. Pennington (2012). Developmental dyslexia, *The Lancet*, vol. 379: 1997-2007.
- Quellette, G. P. & M. Senechal (2008). A window into early literacy: Exploring the cognitive and linguistic underpinnings of invented spelling, *Scientific studies of reading*, vol. 12: 195-219.
- Rapscak, S. Z., P. M. Beeson, M. L. Henry, A. Leyden, E. Kim, K. Rising, S. Andersen & H. Cho (2009). Phonological dyslexia and dysgraphia: cognitive mechanisms and neural substrates, *Cortex*, vol. 45: 575-91.
- Tainturier, M.-J., S. Schiementz & E. C. Leck (2006). Separate orthographic representations for reading and spelling? Evidence from a case of preserved lexical reading and impaired lexical spelling, *Brain and Language*, vol. 99: 31-32.
- Treiman, R. & Kessler, B. (2014). *How children learn to write words*, Oxford University Press.
- Vandermosten, M., B. Boets, H. Poelmans, S. Sunaert, J. Wouter & P. Ghesquière (2012). A tractography study in dyslexia: Neuro-anatomic correlates of orthographic, phonological and speech processing, *Brain*, vol.135: 935-948.
- Viise, N. M., H. C. Richards & M. Pandis (2011). Orthographic depth and spelling acquisition in Estonian and English: A comparison of two diverse alphabetic languages, *Scandinavian Journal of Educational Research*, vol. 455: 425-453.