A Danish, stochastic spelling machine: How to spell a word, when you know part of it

Preben Kihl

”I know part of it”

About the clock. She had learned about the hour hand, but not the minute one

From a certain stage of development Danish spelling is like that

Spelling errors and complex rules

- **Danish spelling is complex and deep: Sounds are spelled in two or more ways and many letters are silent. Could this complexity be usefull?** (Carillo et al., 2012; Furnes & Samuelson, 2010; Hansen, 1967; Molbæk Hansen, 1990, Viise et al., 2011)

- **Read's dogma: Only misspelled words give information about the spelling process, not correctly spelled words** (Read, 1975)

- **For instance**
  ”mig” [maj] MAJ (me), ”lægge” [lɜgə] LEKKE (lay)
  ”begavet” (gifted), ”albaner” (Albanian)

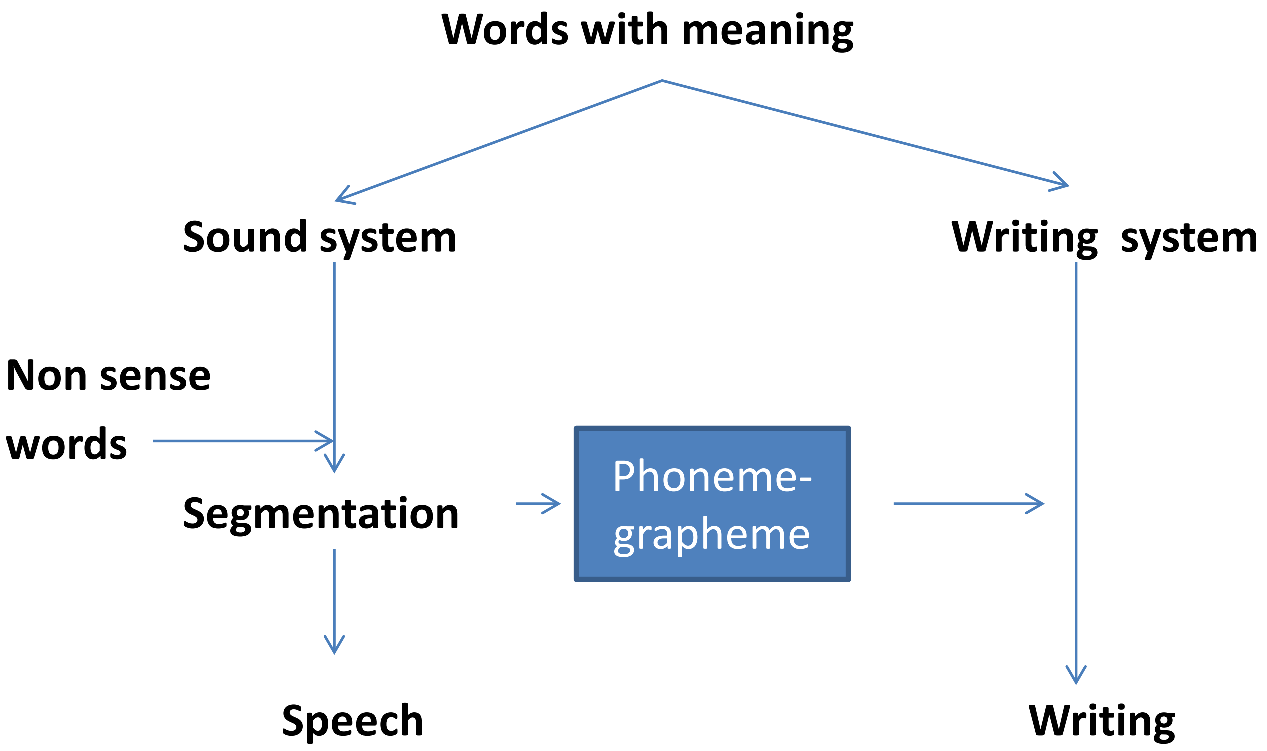
Read is right as

- **You can often interpret the wrong letters in a misspelled word**
- **But what about the correct letters in a misspelled word?**

- **A convenient theoretical frameword for this study: The two-route model**

A two-route model of spelling: The phonological and lexical routes (Ellis, 1987; Rapcsak et. al., 2007; Ziegler et. al., 2008; Vandermosten et al., 2012)

In general the present study is about the left side of the figure, the phonological route. And in particular about the probabilities in the phoneme-grapheme conversion box



Error types analysed in this study

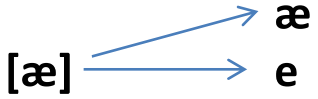
Only phonologically plausible, well ordered errors in accordance with the Danish spelling system. Not chaotic errors from e.g. dyslectics or people with brain damage (Elbro, 1990, 1997; Lennox & Siegel, 1994)

Errors by sounds

Rules

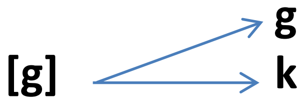
”hest” [hæsd] HÆST (horse)

”gæst” [gæsd] GEST (guest)



”tyk”   [tyg] TYG (fat)

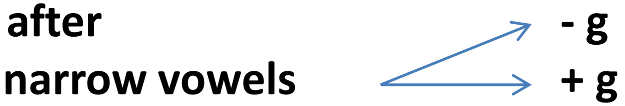
”læg”   [læg] LÆK (put down)



Errors by silent letters

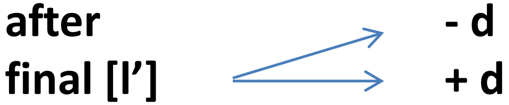
”fugl” [fu’l] FUL   (bird)

”hul” [hu’l/HUGL (hollow)



”fald” [fal’] FAL   (fall)

”hal”   [hal’] HALD (hall)



The problem

- 1. In the typical Danish two-choice situation: How many of each kind of spelling error?
  - E.g. HÆST/GEST (horse/guest): The quantitative relation between incorrect Æ and E letters for the sound [ɜ]?
  - E.g. FAL/HALD (fall/hall): The quantitative relation between incorrectly inserted and deleted D´s ?
- Guesswork? Does letter frequency of occurrence in Danish texts play a part? What about word length and word class? Degree of abstractness? (Kihl, Gregersen & Sterum, 2000)

- 2. Where do the correctly spelled letters in misspelled words come from?

The data

Error frequencies were counted from lists in A. Noesgaard (1945). ”Error types in Danish orthography”. Copenhagen: Fr. Bagges kgl. Hofbogtrykkeri

- In 1940 2 x 10,000 Copenhagen school children in grades 3 to 5 (RI) and 6 to 8 (RII) each spelled 100 words. Accordingly 2 million words. After some data reduction each word was spelled 300 times

- Noesgaard and a bunch of teachers categorized the error types and counted how many errors each type contained

The purpose of my reconstruction

1. Noesgaard and his helpers counted which letters were misspelled as which. I counted how often inferred, underlying sounds were misspelled as one letter or another (or three)

2. To categorize Noesgaard’s data in classic linguistic fashion

3. To analyse the resulting distributions statistically

The counted subsystems

Sounds

Short [ɜ], [ø] og [e]
misspelled as E/Æ, Ø/Y, and E/I

Medial og final stops [b,d,g]
misspelled as B/P, D/T, and K/G

The two components of the [aj]-diphtong misspelled as A/E/I and J/G

Unstressed final [ɔ] misspelled as ER/RE/RER

Silent letters

Silent h before [j] and [v] misspelled as +/- H

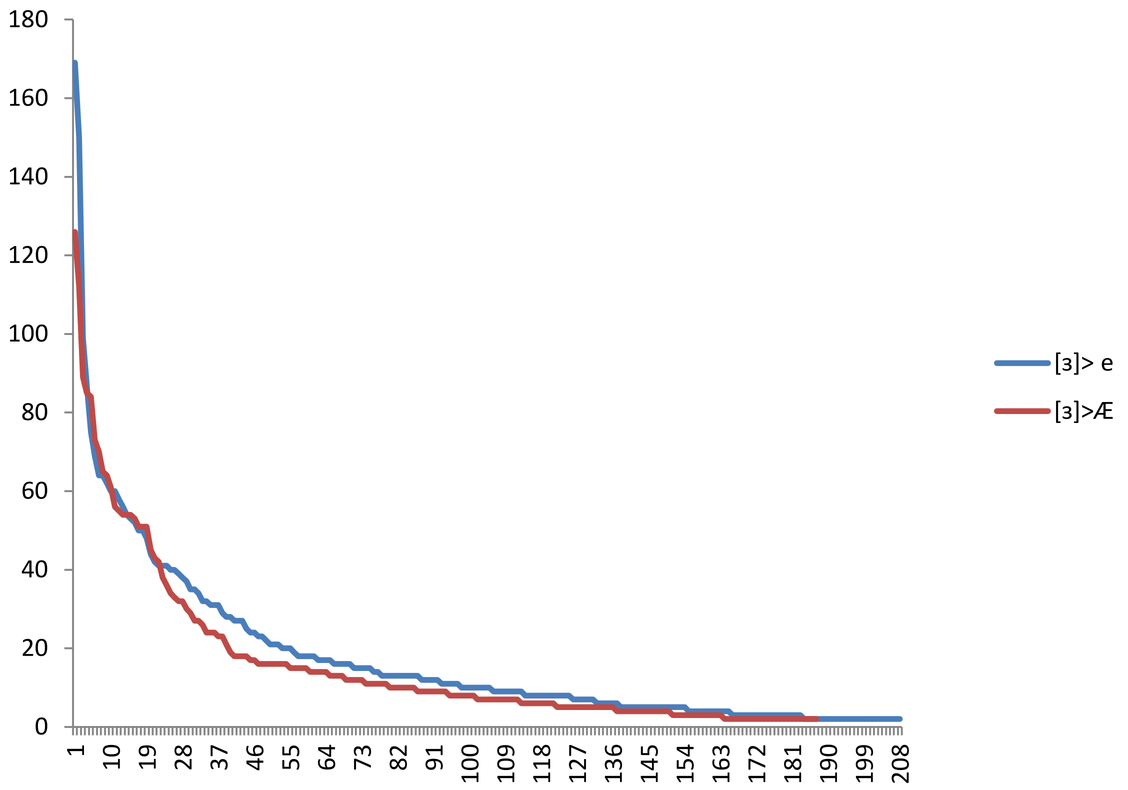
Silent d after final [l], [n], [r] misspelled as +/-d

- **The result of the counts: See next slide**

An unexpected result: 25 out of 29 statistical comparisons showed no significant difference

- Equal numbers of misspelled sounds in pairs, e.g.
- *Vowels*
- ”kys” [kœs] KØS (kiss) /”bøsser” [bœsɔ] BYSSER (guns)
- ”pinde” [penə] PENDE (stick)/”fedt” [fed] FIDT (fat)
- *Stops*
- ”hop” [hɔb] HOB (jump)/ ”snak” [snag] SNAG (talk)/”hat” [had] HAD (hat)
- (and the opposite)
- *Diphthongs*
- ”leg” [laj’] LAJ (play)/”Kaj”[kaj’] KEG (a name)

- Equal numbers of silent letters inserted or left out in pairs, e.g.
- ”hjem” [jɜm’] JEM (home)/”jer” [jɜr] HJER (you)
- ”mord” [mor’] MOR (murder)/”vær” [vɜr’] VÆRD (be)
- ”kald” [kal’] KAL (call)/”bal” [bal’] BALD (ball)



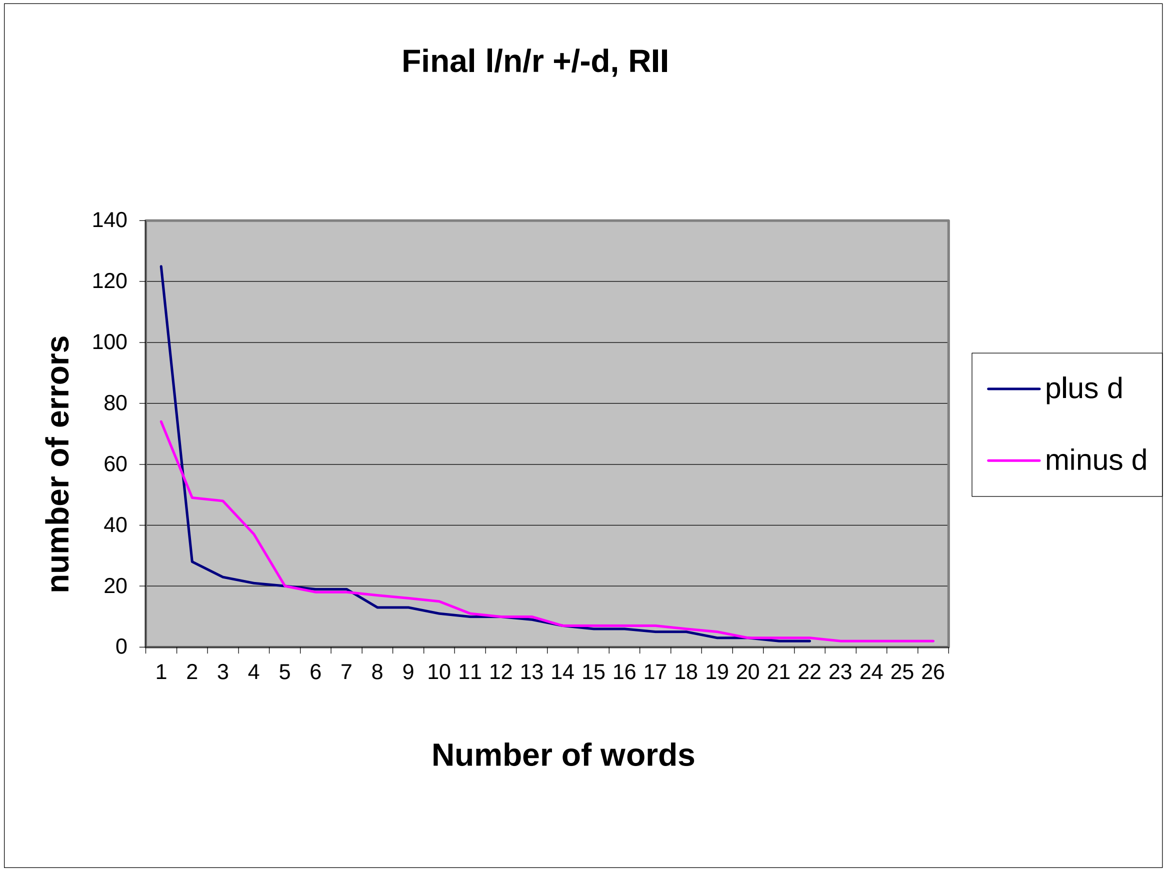
Misspellings of the short vowel [ɜ] vowel as either E or Æ

Subjects 10,000 Copenhagen children, grades 3 to 5. Number of errors on the Y-axis, spelled words on the X-axis. The curves represent 196,500 spelled words (including words with zero errors). 6,063 spelling errors in all.

The numbers behind the curves

(Numbers in the table from both grades 3-5 and 6-8, *no significant difference)

Age-group	Type of error	Words with errors	Number of spelling errors	Average	SD	t-value	Significance (two-tailed)
RI	[æ]>E	189	3042	16.1	20.77		
	[æ]>Æ	206	3621	17.56	22.52	.678	.498*
RII	[æ]>E	124	1400	11.29	12.61		
	[æ]>Æ	136	1887	13.88	20.99	1.22	.225*



Misspelled silent final D's after [l/n/r + glottal stop]

Subjects 10.000 Copenhagen children, grades 6 to 8. The curves represent 14,700 spelled words (words with zero errors not included). 761 spelling errors in all.

Numbers and statistics

no signifikant difference, # signifikant difference

Age group	Type of error	Words with errors	Number of spelling errors	Average	SD	t-value	Signifi-canse (2-tailed)
RII 6.-8. grade	-ld>l	10	103	10.3	6.31		
	-l>ld	11	169	15.36	20.44	-.782	.449*
	-nd>n	10	113	11.3	9.17		
	-n>nd	7	111	15.86	18.96	-.664	.517*
	-rd>r	2	144	72	74.95		
	-r>rd	9	121	13.44	14.48	-2.631	.027#
	In all	22	360	16.36	25.4		
		27	401	14.85	17.58	.237	.814*

Where do the error proportions 50:50 come from? Perhaps they mirror the frequency of occurrence of letters in Danish texts?

- An English dictionary of sound-to-letter frequency of occurrence exists (Rocky, 1973). Used by Baxter & Warrington in a study of a brain damaged man's spelling (1987). But not a Danish dictionary

- After considering the problem for a number of years I have counted the sound-to-letter frequencies of occurrence in the first 1000 most frequent words in Maegaard & Ruus, 1981: ”Frequent words in Danish children's books”. The first 1000 words cover 81.25% af all letter occurrences in the data base.

Sounds and silent letters	Relation between spelling errors	Relation between letter frequencies
RI: short [ɜ]	1,1 : 1	7,63 : 1
RII: Medial og final [g]	1,03 : 1	4,72 : 1
RI: initial [j] +/-H	1,01 : 1	9,13 : 1
RI + RII: final [l,n,r] +/-D	1,41 : 1	3,78 : 1

Representative comparisons between spelling error frequencies from Noesgaard (1945) and sound-to-letter frequencies in Maegaard & Ruus (1981)

It is apparent that the spelling error frequencies do not mirror the frequency of occurrence of letters in Danish children’s books at all. Some comparisons show larger discrepancies.

A half way conclusion

- In the typical Danish spelling choices between letters Noesgaard's subjects choose at random, i.e. the probability of a certain choice is 50:50. Furthermore the choices appear to be context independent/paradigmatic, and are independent of the frequency of occurence of letters in children's books

- A few preliminary correlations between number of spelling errors and word frequence of occurence showed no significant

  connection. There was a weak, significant connection with word length,

  coefficients 0,11 og 0,15

But Noesgaard’s children knew part of it, namely the correct spelling of

- The initial stops [b,d,g,p,t,k]
- The nasals [m,n,ng] (except in French loan words and medially after a short vowel)
- The unvoiced fricatives [f,s,sj,h] (except in foreign words and medially after a short vowel)
- The voiced fricatives [v,j,r]

- The short, long and ”stød” vowels [i,y,ö,u,o,a, schwa]
- The remaining long and ”stød” vowels (except [ɜ,å])

Why? What is easy and what is difficult?

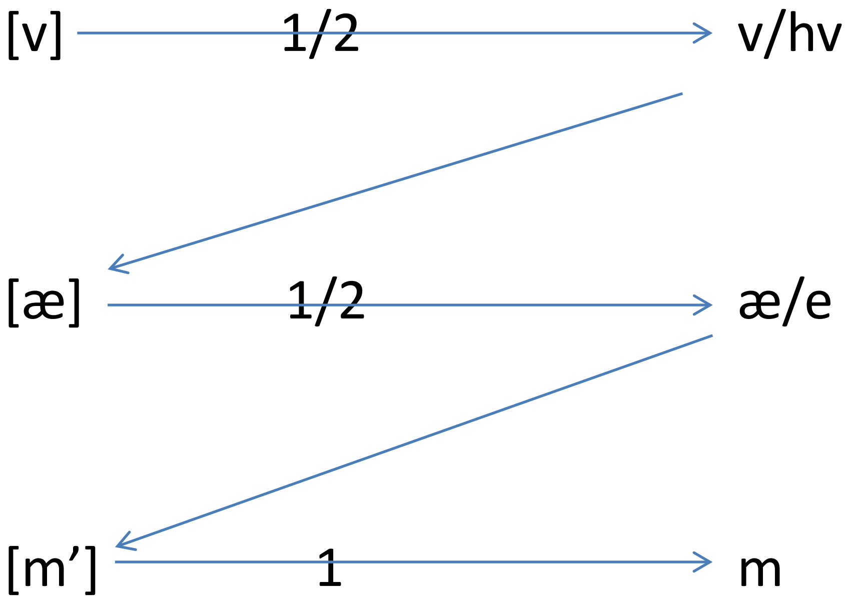
- Danish one-one relations between sound and letter are spelled correctly because they are easy to spell (obviously circular)

- The statement should, however, be seen in the light of the one-two relations that obviously lifelong are not easy to learn

- Accordingly the probabilities are in place: The Danish, stochastic spelling machine can be presented

An outline of a non-deterministic spelling machine that after segmentation of an unknown word spells from left to right. Probabilities on straight arrows pointing right, the oblique arrows indicate back-coupling

”hvem” [vзm’] (who)



[v] 1/2 v/hv

[æ] 1/2 æ/e

[m’] 1 m

A stochastic matrix of transition probabilites. The spelling of ”hvem” (who)

Operands/ Transforms	[v]	[æ]	[m’]
v	½	0	0
hv	½	0	0
æ	0	½	0
e	0	½	0
m	0	0	1

Put a word, e.g. ”skole” (school), into the machine and watch what happens

Operands/ Transforms	[sg]	[o:]	[l]	[ə]
sk	1	0	0	0
o	0	1	0	0
l	0	0	1	0
e	0	0	0	1

Operands/ Transforms	[v]	[œ]	[l]*	[ɔ]
v	½	0	0	0
hv	½	0	0	o
ø	0	½	0	0
y	0	½	0	0
l	0	0	1/3	0
ll	0	0	1/3	0
ld	0	0	1/3	0
rer	0	0	0	½
re	0	o	0	½

The spelling of a Danish nonsens word [vœlɔ]

Notice the wealth of spelling possibilities that a very simple sound structure presents

*[l] is misspelled in the three ways shown above, but the probabilities are unknown

The spelling of correct and incorrect letters in misspelled words

It is well known that both misspelled words and nonsense words must be segmented before they are spelled (e.g. Holligan & Johnston, 2005; Rapzack et al., 2009)

The present study has shown that correct and incorrect letters in a misspelled word are spelled in the same way after segmentation. The difference between correct and incorrect letters is the attached probabilities, nothing else. All sounds in unknown words (with or without meaning) are spelled, only the probabilities are different.

A deduction from the theory: Potential spelling errors and the size of the spelling vocabulary processed by the lexical route

- The word ”skal” [sgal] (must) was misspelled as SKALD 74 times in the Noesgaard study (18 other errors will be ignored). The present study has shown, that approximately 74 other children must also have thought about a D-letter, but rejected it.
- Accordingly 148 children did not know the spelling of the word ”skal”. Half of them had luck, the other half not.
- You may remember that every word in Noesgaard’s study was spelled 300 times. Accordingly 152 children, approximately 50%, had learned and knew the spelling of the word ”skal” by heart . These words did not pass through the spelling machine on the phonological route, but in theory moved through the lexical route.

Continued: Danish words very often present several error possibilities

- The following probabilites must apply, when the spelling of a word is unknown (given a certain level of development and some experience with reading and spelling):

- 0 error possibilities in a word 100% **chance of correct spelling**
- 1 error possibility in a word 50 % ”
- 2 error possibilities in a word 25 % ”
- 3 error possibilities in a word 12½ % ”
- 4 error possibilities in a word 6 ¼ % ”

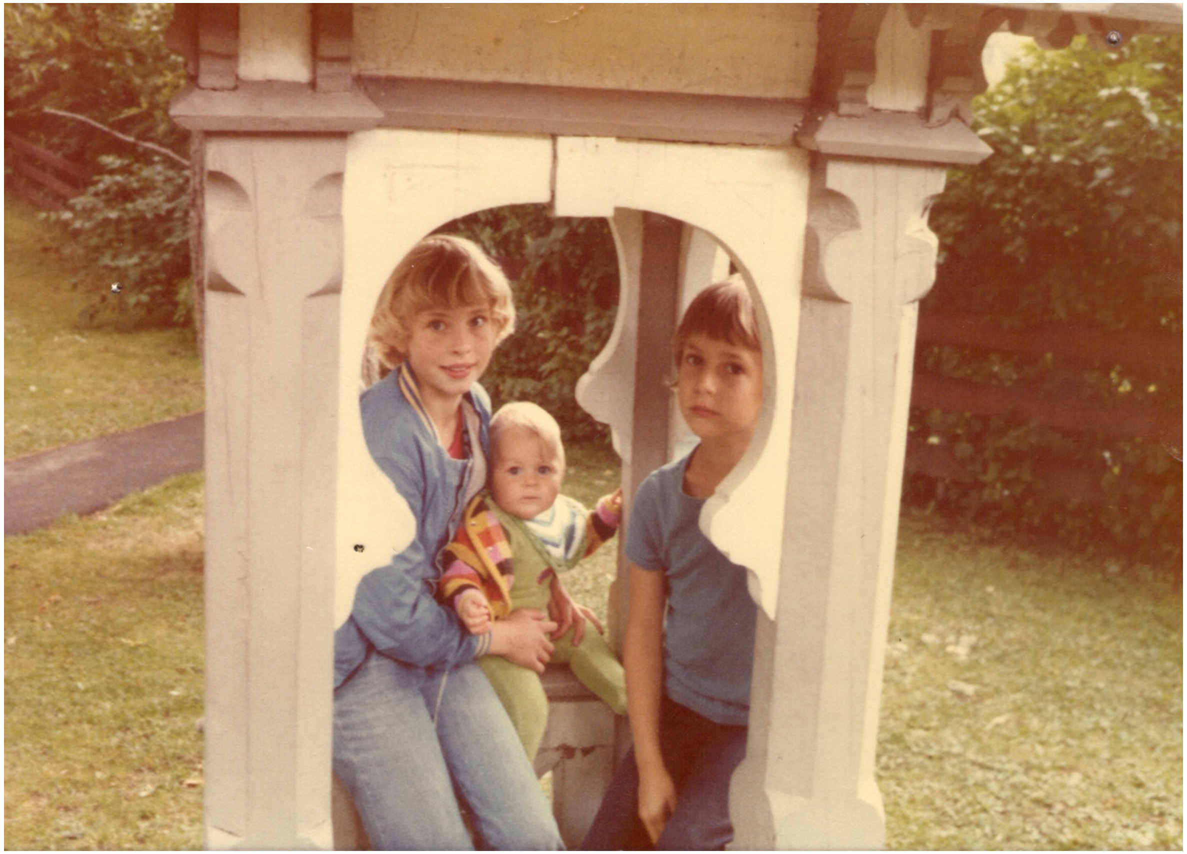
A sketch of a test that measures the size of the lexical spelling vocabulary

- Calculate word by word in a given word material the chance of accidental correct spelling (base the calculations on Noesgaard and this presentation)
- When a subject has done the test, calculate how many of the correctly spelled word that could have been spelled correctly accidentally.
- Add this number to the number of misspelled words and subtract that from the total number of words in the test.
- The resulting number is a measure of the test-subject's lexical spelling vocabulary. Whatever that is.

  (Ehri, 2009; Martinet et al., 2004; Tanturier et al., 2006)

- Thank you for your attention

Actual and potential Copenhagen school children



References

Baxter, D.M., & Warrington, E.K (1987). Transcoding sound to spelling: Single or multiple sound unit correspondences? *Cortex,* 23, 11-18.

Carillo, M.S., Alegria, J, & Marin, J. (2012). On the acquisition of some basic word spelling mechanisms in a deep (french) and a shallow (spanish) system. *Reading and writing*, online.

Ehri, L.C. (2009). Learning to read words: Theory, Findings, and Issues. *Scientific studies of Reading,* Vol. 9, 167-188.

Elbro, C. (1990). *Differences in dyslexia. A study of reading strategies and deficits in a linguistic perspective*. København: Munksgaard.

Elbro, C. (2007). *Læsevanskeligheder*. København: Gyldendal.

Ellis, A.W. (1987). Modelling the spelling process. G. Denes, C. Semenza, P. Bisiacchi & E. Andreewsky (ed.), *Perspectives in cognitive Neuropsychology.* London: LEA. 189-211*.*

Furnes, B., & Samuelsson, S. (2010). Predicting reading and spelling difficulties in transparent and opaque orthographies: a comparison between Scandinavian and US/Australian children. *Dyslexia*, Vol. 16, 119-142.

Hansen, Aa. *Moderne Dansk I*. København: Grafisk Forlag, 1967.

Kihl, P., Gregersen, K., & Sterum, N. (2000). Hans Christian Andersen's spelling and syntax: Allegations of specific dyslexia are unfounded. *Journal of learning disabilities*, Vol. 33 (3), 506-519.

Lennox, C., & Siegel, L.S. (1994). The role of phonological and orthographic processes in learning to spell. G.D.A. Brown & N. C.Ellis (ed.), *Handbook of spelling*. Chichester: Wiley &Sons. 93-110.

Maegaard, B., & Ruus, H. (1981). *Hyppige ord i danske børnebøger*. København: Gyldendal.

Martinet, C., Valdois, S., & Fayol, M. (2004). Lexical orthographic knowledge develops from the beginning of literacy acquisition. *Cognition,* Vol. 91, B11-B22.

Molbæk Hansen, P. (1990). Talende systemer: anvendelser og lingvistiske udfordringer. I Engberg Petersen m.fl. (red.). *Anvendt sprogvidenskab*. København Universitet: Museum Tuscalanums Forlag, 15-35.

Noesgaard, A. (1945). *Fejltyper i dansk retskrivning*. København: Fr. Bagges kgl. Hofbogtrykkeri.

Rapcsak, S.Z., Henry, M.L., Teague, S.L., Carnahan, S.D., & Beesru, P.M. (2007). Do dual-route models accurately predict reading and spelling performance in individuals with acquired alexia and agraphia? *Neuropsychologia*, Vol. 45, 2519-2524.

Molbæk Hansen, P. (1990). Talende systemer: anvendelser og lingvistiske udfordringer. I Engberg Petersen m.fl. (red.). *Anvendt sprogvidenskab*. København Universitet: Museum Tuscalanums Forlag, 15-35.

Noesgaard, A. (1945). *Fejltyper i dansk retskrivning*. København: Fr. Bagges kgl. Hofbogtrykkeri.

Rapcsak, S.Z., Henry, M.L., Teague, S.L., Carnahan, S.D., & Beesru, P.M. (2007). Do dual-route models accurately predict reading and spelling performance in individuals with acquired alexia and agraphia? *Neuropsychologia*, Vol. 45, 2519-2524.

Rapscak, S.Z., Beeson, P.M., Henry, M.L., Leyden, A., Kim, E., Rising, K., Andersen, S., & Cho, H. (2009). Phonological dyslexia and dysgraphia: cognitive mechanisms and neural substrates. *Cortex,* Vol. 45, 575-91.

Read, C. (1975). *Children's categorizations of speech sounds in English.* Urbana Ill.: NTCE, 1975.

Rockey, D. (1973). *Phonetic lexicon*. London: Heyden and son.

Tainturier, M.-J., Schiementz, S., & Leck, E.C. (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.

Vandermosten, M., Boets, B., Poelmans, H., Sunaert, S., Wouters, J., & Ghesquière, P. (2012). A tractography study in dyslexia: neuroanatomic correlates of orthographic, phonological and speech processing. *Brain*, Vol. 135, 935-948.

Ziegler, J.C., Castel, C., Pech-Georgel, C., George, F., Alario, X-F., & Perry, C. (2008). Developmental dyslexia and the dual route models of reading: Simulation individual differences and subtypes. *Cognition,* Vol. 107, 151-178.

Viise, N.M., Richards, H.C., & Pandis, Meeli (2011). Ortographic dept and spelling acquisition in Estonian and English: A comparison of two diverse alphabetic languages. *Scandinavian Journal of Educational Research*, Vol. 55, 425-453.