Jón Friðrik Daðason, Kristín Bjarnadóttir og Kristján Rúnarsson

Skrambans villurnar:
Villugreining á tölvutækum textum

28. Rask-ráðstefnan um íslenskt mál og almenna málfræði
25. janúar 2014

STOFNUN ÁRNA MAGNÚSSONAR
Í ÍSLENSKUM FRÆÐUM

Efnið: Niðurstöður og flokkun á villum úr Skramba

• Skrambi er leiðréttingarforrit. Upplýsingar um verkefnið sem villugreiningin er unnin úr eru á dreifiblaði.*

• Villuflokkun Skramba: Orðleysisvillur og samhengisháðar villur
• Almenn flokkun á villum: Stafsetningarvillur, ritvillur, málvillur
• Sértækar villur í tölvutækum textum

• Samanburður við flokkun Baldurs Sigurðssonar og Steingríms Þórðarsonar á stafsetningarvillum á grunnskólaprófi í *Íslensku máli* 1987.

• Villusafnið: 5.000 orðleysisvillur og algengustu samhengisháðar villur. Tíðnitölur eru úr vefgögnum og þær eru ekki jafntraustar og tölur úr handyfirförnum gögnum.

*Texti dreifiblaðsins er aftast í þessu skjali.

Orðleysisvillur og samhengisháðar villur

Orðleysisvilla er orð sem ekki er til í tungumálinu, þ.e. orðmynd sem aldrei getur verið rétt skrifað íslenskt orð:

kanske, fleirri, uppgvötva, fyrifram, tréinu, alkahól

Samhengisháð villa er orð sem er til í málinu en er rangt í samhenginu sem það birtist í:

***Himininn** er blár*
*Ég horfi upp í **himinninn***
*Hún bað hann um að **sína** sér bókina **sýna***

Hlutfall samhengisháðra villna í íslenskum textum: 68%
(skv. greiningu Kristjáns Rúnarssonar á leiðréttum framhaldsskólaritgerðum 2011, sjá dreifiblað).
Til samanburðar er hlutfallið í ensku 25-40% (Kukich 1992).

Meginskipting orðleysisvillna

Stafsetningarvillur: Brot á stafsetningarreglum

Klaufavillur: Ritvillur, ásláttarvillur, stafavíxl ...

Þriðja gerðin: Sértækar villur í tölvutækum textum, oft vegna vandræða með stafasett: *thessi, thurftum* ...

Þriðja gerðin verður ekki skoðuð hér og ekki Sms-stafsetning heldur.

Óviðráðanlegu villurnar:

Foreldrarnir setja eins konar stall sem börn þeirra lifa eftir og vilja fylla sömu skó.

Stafsetning: Helstu gerðir orðleysisvillna

	Villur	*Dæmi*
Orðbundin villa	226	526.367 *einhvað*
Eitt orð eða tvö	159	471.410 *ennþá*
Ein- eða tvíritað samhljóð	310	65.099 *kanske*
Villa í samsetningu	500	64.227 *Oddson*
Eitt eða tvö n í bakstöðu	232	51.665 *allann*
Villa í beygingarendingu	326	43.620 *vinnuni*
Y	232	39.333 *partý*
J-reglan	63	16.883 *nýjir*
Beygingarvillur	68	8.348 *fæturnar*
Z	135	6.525 *áherzla*
E f. ei og ei f. e	27	3.268 *meigi*
Sérhlj. á undan ng/nk	58	3.260 *einginn*

Stafsetning: Helstu gerðir orðleysisvillna, frh.

	Villur	*Dæmi*
Je fyrir é	28	1.213 *Jésú, ské*
Hv og kv	12	1.169 *áhvað*
Bandstrik	31	696 *suðurafríska*
Stafsetningarvillur alls	2.081	1.259.563

Orðbundnar villur

	Villur	*Dæmi*	
einhver	51	24.098	*einhvað, eikkað ...*
kven-	13	749	*kvennkyns, kvennfólk ...*
nokkur	10	1.441	*nokkurar*
svolítill	9	1.097	*svolitið*
hvernig	9	1.766	*hverning*
þessi	8	1.101	*þettað*
svoleiðis	*8*	1.200	*sollis, soleiðis*
fyrir	*8*	921	*fyirir, fyriri, fyrrir*
einkunn	*7*	870	*einkunina*
herbergi	*6*	196	*herberja, herbegi*
einmana	*6*	1.101	*einmannalegt*
...			
tvímælalaus	*2*	77	*tvímannalaust*

Klaufavillur, innsláttarvillur

	Villur	*Dæmi*	
Einn staf vantar	729	50.334	*aðalega*
Íslenska stafi vantar	749	26.742	*adeins**
Brodd vantar	272	20.659	*júni*
Einum staf ofaukið	218	16.193	*Afghanistan*
Stafavíxl	168	11.883	*gengdi*
Broddi ofaukið	31	4.538	*ílla*
Ásláttarvillur	16	1.486	*mað, e´g*
Klaufavillur alls:	2.183	131.724	

* Líka einkenni á tölvutækum textum.

Stafsetningarvillur eru algengar, ritvillur eru hending

	Villur	*Dæmi*	*Meðaldæmafjöldi*
Stafsetningarvillur alls:	2.081	1.259.563	605
Klaufavillur alls:	2.183	131.724	60

Textaflokkar í MÍM:
Tíðni 5.000 algengustu orðleysisvillna

	Villur	*Lesmálsorð*	
blogg	6.791	2.184.474	**0,31%**
ritgerðir	568	193.095	0,29%
tölvupóstur	173	134.201	0,13%
vefmiðlar	262	267.096	0,10%
bækur	5.390	6.746.217	0,08%
ruv	172	282.162	0,06%
Vísindavefurinn	928	2.041.013	0,05%
ráðuneyti	708	1.777.638	0,04%
dómar	179	968.026	0,02%
lög	45	438.193	**0,01%**
Alþingi	6	331.256	0,00%
...			
Samtals	22.941	27.527.815	0,08%

Villukóðar Kristjáns

Málnotkun:	s: orðasamband, o/r: rangt orð, u: ummyndun á réttu orði, z: orðaröð
Málfræði:	2: tala, k: kyn, f: fall, t: tíð, á: ákveðni 3: stig, h: háttur (vh/fh, nh/lhþt), ö: persóna g: beyging (rangur flokkur, stofn misskilinn ...) c/m: samræmi, p: samsetningarvilla, l: hliðstætt/sérstætt,
Stafsetning:	v: stafsetningarvilla, 1: orðaskil, A: stór/lítill stafur i: innsláttarvilla, ó: ósamhengisháð (orðmynd ekki til)
Annað:	!: skoða nánar, -: stakt orð eða samsetningarliður (m/tvítekningu) e: óþörf/ólógísk endurtekning w: erlent orð leiðrétt x: kolvitlaust / óleiðréttanlegt

Handunnin leiðrétting, ekki á færi neins forrits!

Samhengisháðar villur

Íslenskar orðmyndir sem eru á röngum stað í texta:

- Orðbundið: *leyti/leiti* (2 orð, sami orðflokkur)
- Orð, orðflokkur: *sína/sýna* (2 orð, ekki sami orðflokkur)
- Röng greining: *himinninn/himininn* (1 orð, 2 föll)

Tölur um samhengisháðar villur eru ekki fyrirliggjandi
en skv. útreikningum JFD/KR gætu þær verið 68% af villum.

Skrambi leiðréttir samhengisháðar villur með því að reikna
út líkindi fyrir rétt orð úr **vafaorðamengi** í mörkuðum texta.

Vafaorðamengi og tíðnitölur

1. *sína* (62.914) / *sýna* (25.988)
2. *komin* (31.013) / *kominn* (21.446)
3. *búinn* (15.219) / *búin* (15.036)

13. *nýju* (28.082) / *níu* (6.343)

32. *breytt* (31.056) / *breitt* (2.632)

904. girða (773) / gyrða (4)

--. *annarra (?) / annara (?)*

annara í orðasambandinu ***e-m er annara um e-ð***,
annars alltaf ***annarra****!*

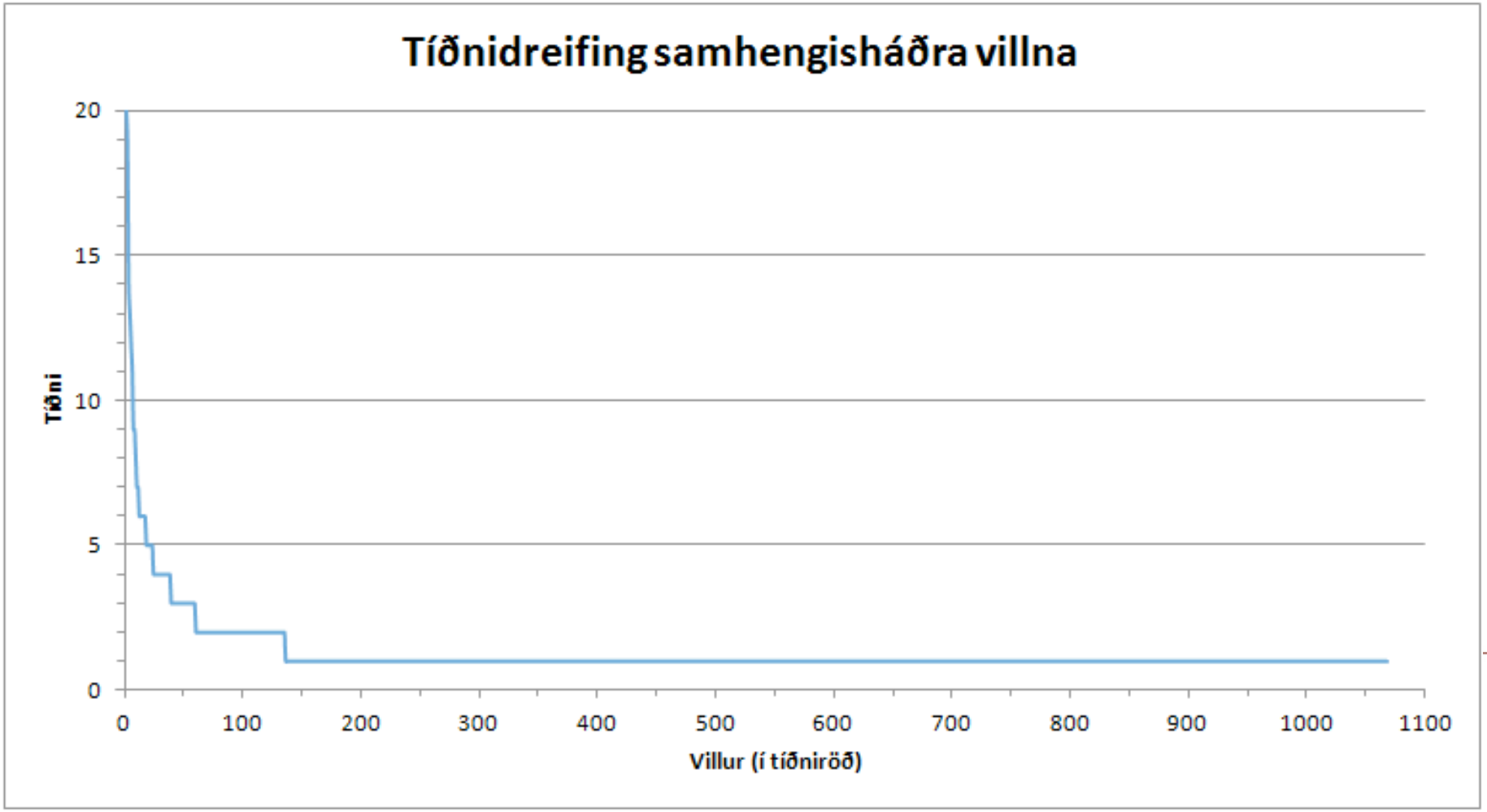
Lágmarkspör úr *Hjali* (hljóðritun og stafsetning),
tíðnitölur úr *Íslenskum orðasjóði.*

Vafaorðamengi úr ritgerðum

20 {þau, þeir}
19 {er, sé}
14 {er, eru}
13 {af, að}
12 {eitthvað, eitthvert}
11 {**enn, en**}
9 {eru, séu}
8 {að, af}
7 {finnst, finnast}
7 {hafði, hefði}
6 {á, eigi}
6 {eru, er}

6 {því, það}
6 {hvort, hvert}
6 {**leiti, leyti**}
6 {**líkur, lýkur**}
5 {í, á}
5 {inn, inni}
5 {**tíman, tímann**}
5 {**allavega, alla vega**}
5 {**bíður, býður**}
5 {**engin, enginn**}
4 {hafa, hafi}
4 {lang flestir, langflestir}

Vafaorðamengin eru í tíðniröð.
Stafsetningarvillur feitletraðar.


Tíðnidreifing samhengisháðra villna
Tíðni
20
15
10
5
0
0
100
200
300
400
500
600
700
800
900
1000
1100
Villur (í tíðniröð)

Baldur Sigurðsson og Steingrímur Þórðarson 1987

Ef menn vissu nákvæmlega hvernig stafsetning lærðist væri ugglaust hægara að kenna hana. (*Íslenskt mál* 9:7--22)

- Úrtakið í könnun BS og SÞ á grunnskólaprófi í **stafsetningu** 1984 voru 989 nemendur.
- M.a. greining á tegundum villna og tíðni
- Stig í stafsetningarnámi:
    - Hljóðgreining (hljóðréttar og hljóðrangar villur (*fæ**k**unar*))
    - Sjónminni / rithefð (sérstaklega í algengum orðum (*f**y**rir*)
    - Stofn orða / orðskilningur (*gru**nn**t, l**ey**nist*)
    - Beygingarfræði (t.d. **n/nn** í enda orðs)
    - Minnisatriði (*anna**rr**a, óþu**rr**kar, stórfe**ll**drar*)

Síðari þættirnir þrír falla undir það sem ætla má að nemandi beiti meðvitað m.a. fyrir tilstilli stafsetningarkennslu.

Allar þessar gerðir af villum koma líka fram í greiningu okkar**.**

Hvaða atriði í stafsetningu er erfiðust?

„Okkar svar við spurningunni er því að erfiðustu atriðin séu hvorki Y né N heldur **hrein minnisatriði**. Sé orðið ókunnugt, séu til samhljóma orðmyndir, sé langsótt að skýra rithátt, þá er orðið erfitt. “ (BS og SÞ 1987)

Sbr. villutíðni í greiningu okkar á orðleysisvillum:

1. Orðbundnar villur (þ.e. minnisatriði)
2. Eitt orð eða tvö
3. Ein- eða tvíritað samhljóð
4. Villa í samsetningu
5. Eitt eða tvö n í bakstöðu
6. Villa í beygingarendingu
7. Y

Meirihluti villna er samhengisháður!

Í hverju er Skrambi bestur? Í því sem fólk gerir oft!

Lokaorð

• Skrambi er bestur í því að leiðrétta villur sem hann hefur nægileg gögn um, þ.e. villur sem fólk gerir oft. Samkvæmt villutölum okkar eru það einmitt erfiðustu villurnar, minnisatriðin.
• Skrambi dugar því vel til að grynnka verulega á villum.
• Skrambi ætti að nýtast í stafsetningarkennslu með því að benda fólki jafnóðum á vafaatriði og leggja til leiðréttingar.
• Við þetta gæti sparast tími sem nýta mætti í umfjöllun um mál, stíl og hugsun, sbr. dæmið:

Foreldrarnir setja eins konar stall sem börn þeirra lifa eftir og vilja fylla sömu skó.

Takk fyrir áheyrnina.

Jón Friðrik: jfd1@hi.is
Kristín B: kristinb@hi.is
Kristján R: krr1@hi.is

STOFNUN ÁRNA MAGNÚSSONAR
Í ÍSLENSKUM FRÆÐUM

51 vitlaus riháttur fyrir *einhver*

einhvað 6.176 eithvað 4.945 eitthver 2.808 eikkað 2.244
eitthvern 1.146 eitthverja 755 eitthverjum 570 eitthverjar 416
eikkvað 340 eitthverjir 286 eikker 280 eittvað 246 eithver 246
ekkað 234 eitthað 192 einnhver 166 enhver 158 eikkva 154
eitthverri 150 eihverjar 131 einvhern 124 etthvað 121
einhverir 118 eikkern 113 einver 111 einvherjum 102 eihver 102
einnhvern 98 eikað 98 einvherju 94 einkvað 87 einkver 83
eikkerja 82 einhveri 81 eithvern 79 ikkað 77 einvern 74 eithvert 73
eithverja 71 einhvur 70 eikkur 70 eihvað 68 einvherjir 67
einverjum 67 eiithvað 65 einvherjar 62 einhven 61 eiker 61
enhvað 58 itthvað 13 einhverni 5

[Dreifiblað:]Skrambans villurnar:
Villugreining á tölvutækum textum

Jón Friðrik Daðason, Kristín Bjarnadóttir og Kristján Rúnarsson
Stofnun Árna Magnússonar í íslenskum fræðum
jfd1@hi.is, kristinb@hi.is, krr1@hi.is

Um verkefnin sem villugreiningin er unnin úr

Sumarið 2011 var unnið að villugreiningu á orðfræðisviði Stofnunar Árna Magnússonar í íslenskum fræðum með það í huga að undirbúa gerð leiðréttingarforrits fyrir samhengisháða stafsetningarleiðréttingu. Vinnan fól í sér handunna villugreiningu en síðan voru þróaðar vélrænar greiningar- og leiðréttingaraðferðir. Verkið var unnið af Jóni Friðriki Daðasyni og Kristjáni Rúnarssyni með styrk frá Nýsköpunarsjóði námsmanna og Vinnumálastofnun. Umsjónarmenn Nýsköpunarsjóðsverkefnisins voru Sven Þ. Sigurðsson, prófessor í tölvunarfræði við HÍ, og Kristín Bjarnadóttir, rannsóknarlektor hjá Stofnun Árna Magnússonar í íslenskum fræðum. Verkefnisstjóri var Kristín Bjarnadóttir.

Í handleiðréttingunni fólst lestur og leiðrétting á ritgerðum framhaldsskólanema og voru villurnar flokkaðar jafnóðum og þeim skipt í samhengisháðar villur og orðleysisvillur.

Leiðréttingarforritið Skrambi

Skrambi er nýtt stafsetningarleiðréttingarforrit sem er afrakstur af verkefnum sem unnin hafa verið á Árnastofnun frá árinu 2010, þ.e. Leiðrétting ljóslesinna texta (2010), Samhengisháð villuleiðrétting (2011) og Fjölnir fyrir hvern mann (2012), auk meistaraverkefnis Jóns Friðriks, sjá skemman.is/en/item/view/1946/12085.
Skrambi er til ýmissa nota gagnlegur og leiðréttir ljóslesinn texta, færir eldri texta til nútímamáls og leiðréttir stafsetningu, bæði nútímamálstexta og eldri texta.

Skrambi hlaut hagnýtingarverðlaun Háskóla Íslands árið 2012.

frh. ->

Villusafnið

Orðleysisvillum var safnað úr Íslenskum orðasjóði eftir villulistum sem til urðu við vinnuna við Skramba. Í listanum er gefin tíðni 5.000 algengustu villnanna.

Heimildir

Baldur Sigurðsson og Steingrímur Þórðarson. 1987. Hvernig geta börn lært stafsetningu? *Íslenskt mál* 9:7-21.

Jón Friðrik Daðason. 2011. Hönnun hugbúnaðar fyrir samhengisháða stafsetningarleiðréttingu. Lokaskýrsla til nýsköpunarsjóðs námsmanna.

Kukich, K. 1992. *Techniques for Automatically Correcting Words in Text. ACM Computing Surveys,* 24 (4), 377-439.

Mörkuð íslensk málheild, sjá mim.hi.is

Íslenskur orðasjóður, sjá http://wortschatz.uni-leipzig.de/ws_isl/

STOFNUN ÁRNA MAGNÚSSONAR
Í ÍSLENSKUM FRÆÐUM