Leiðrétting á ljóslesnum textum

Kristín Bjarnadóttir og Jón Friðrik Daðason
Orðfræðisvið SÁ

Fyrirlestraröð Máltækniseturs
2. nóvember 2010

STOFNUN ÁRNA MAGNÚSSONAR
Í ÍSLENSKUM FRÆÐUM

Verkefnið *Leiðrétting á ljóslesnum textum*

- Hluti af átaki Vinnumálastofnunar: *856 störf*

- Felst í því að þróa aðferðir og hugbúnað til leiðréttingar á skönnuðum íslenskum textum frá 19. öld

- Textarnir eru frá Landsbókasafni-Háskólabókasafni, af timarit.is

- Þátttakendur: Jón Friðrik Daðason, Kristján Rúnarsson,
  Kristín Bjarnadóttir,
  Sigrún Helgadóttir, Ásta Svavarsdóttir

Tímarit.is, Landsbókasafn-Háskólabókasafn

Stafrænt bókasafn sem veitir aðgang að milljónum myndaðra blaðsíðna á stafrænu formi af þeim prentaða menningararfi sem varðveittur er í blöðum og tímaritum frá Færeyjum, Grænlandi og Íslandi, sjá timarit.is: Um vefinn ...

Heildarfjöldi myndaðra blaðsíðna: 3.467.276
Heildarfjöldi ljóslesinna blaðsíðna: 3.420.885

ÍBN 19s-20f, Stofnun Árna Magnússonar í íslenskum fræðum, orðfræðisvið

Safn blaða- og nytjatexta frá síðari hluta 19. aldar til fyrri hluta 20. aldar, ætlað til rannsókna á málfari. Textarnir eru úrtak af timarit.is.

625 textar, 4.900.884 lesmálsorð

STOFNUN ÁRNA MAGNÚSSONAR
Í ÍSLENSKUM FRÆÐUM

Markmiðið

- að þróa aðferðir og hugbúnað til leiðréttingar á skönnuðum íslenskum textum frá 19. öld

Aðferð

- Grunnhugmyndin er að bera saman orðmyndir úr textunum og úr gögnum orðfræðisviðs Stofnunar Árna Magnússonar í íslenskum fræðum.
- Útkoman er listi um óþekktar orðmyndir, þ.e. hugsanlegar villur.
- Leiðrétting á “villunum”, með aðferðum sem JFD lýsir hér á eftir, m.a. með því að nota upplýsingar um mögulegar villur og tíðni þeirra.
- Hluti textanna var handleiðréttur og er notaður
    - sem prófunarsafn
    - í vörpunartöflur (villa/leiðrétting)

Umfang ÍBN 19s-20f

- Handleiðréttir textar: 24 textar, 151.679 lesmálsorð, 280 bls.
- Keyrslutextar: 625 textar, 4.900.884 lesmálsorð

Sýnishorn: Óleiðréttur texti

http://timarit.is/files/9797185.pdf#navpanes=1/9797185.pdf

*) Sigurður Gunnarsson nefnir það (í Norðanfara, 15. árg., 1876, bls. 73), að uppsprettur J'ökulsár séu vestanvert við lvvei'kfjöll en ekki víð Kistufell og segist liafa farið par yfir aðalíarveg arranar með Gunnlaugssen, en hklega lietir árrenslið fallið úr á uppdrættinum pegar hann var prent- aður í Kaupmannahöíh.
**) Mývetningar kalla eldgígj hveri og hefir það stund- um vahlið misslulnmgi; eptir að Mývetniugar fóru til Ðyngjutjalla 1875 og höfðu seð gígiua par, en kallað pa í lýsmgingunni hven, héldu menn annarsstaðar að það væri vatnshverir og í sunnlenzkum og útlendum blöðum var talað um hve merkiiegt það væri að fundinn væri stærri hver en Geysir.

Norðlingur, 7. maí 1881

Sýnishorn: Handleiðréttur texti

*) Sigurður Gunnarsson nefnir það (í Norðanfara, 15. árg., 1876, bls. 73), að uppsprettur Jökulsár séu vestanvert við Kverkfjöll en ekki víð<prentv við /> Kistufell og segist hafa farið þar yfir aðalfarveg árinnar með Gunnlaugssen, en liklega<stafsvafi líklega /> hefir árrenslið fallið úr á uppdrættinum þegar hann var prent-aður í Kaupmannahöfn.
**) Mývetningar kalla eldgígi hveri og hefir það stund-
um valdið misskilningi; eptir að Mývetningar fóru til Dyngjufjalla 1875 og höfðu séð gígina þar, en kallað þá í lýsingingunni<prentv lýsingunni /> hveri, héldu menn annarsstaðar að það væri vatnshverir og í sunnlenzkum og útlendum blöðum
var talað um hve merkilegt það væri að fundinn væri stærri hver en Geysir.

Norðlingur, 7. maí 1881

Leiðrétt af Kristjáni Rúnarssyni

Sýnishorn: Vélleiðréttur texti

*) Sigurður Gunnarsson nefnir það (í Norðanfara, 15. árg., 1876, bls. 73), að uppsprettur Jökulsár séu vestanvert við lvvei'kfjöll en ekki víð Kistufell og segist hafa farið þar yfir aðalfarveg arraðar með Gunnlaugsson, en hklegt litir árrennslið fallið úr á uppdrættinum þegar hann var prentaður í Kaupmannahöfn.
**) Mývetningar kalla eldgíg hveri og hefir það stundum valið misslulnmgi; eptir að Mývetningar fóru til Dyngjufjalla 1875 og höfðu séð gígina þar, en kallað þa í lýsingunni hve, héldu menn annarsstaðar að það væri vatnshverir og í sunnlenskum og útlendum blöðum var talað um hve merkilegt það væri að fundinn væri stærri hver en Geysir.

Leiðrétt
Óleiðrétt
Vitlaust leiðrétt

Athugið! Þessi texti er óvenju slæmur!

Samanburður á sýnishornunum

ORG: lvvei'kfjöll en ekki víð Kistufell og segist liafa farið par
JFD: lvvei'kfjöll en ekki víð Kistufell og segist hafa farið þar
KR: Kverkfjöll en ekki víð<prentv við /> Kistufell og segist hafa farið þar

ORG: pa í lýsmgingunni hven, héldu menn annarsstaðar að það
JFD: þa í **lýsingunni** hve, héldu menn annarsstaðar að það
KR: þá í lýsingingunni<prentv lýsingunni /> hveri, héldu menn annarsstaðar að það

ORG: Ðyngjutjalla 1875 og höfðu seð gígiua par, en kallað
JFD: Dyngjufjalla 1875 og höfðu séð gígina þar, en kallað
KR: Dyngjufjalla 1875 og höfðu séð gígina þar, en kallað

Leiðrétt
Vitlaust
Vitlaust leiðrétt

Norðlingur (7. maí 1881)

- 1602/2013 rétt orð (**79,58%**)
  - 62 þeirra vantar í listann og er því breytt fyrir mistök (3,9%)

- 411/2013 vitlaus orð (**20,42%**)
  - 272 þeirra voru leiðrétt (66,2%)
  - 105 þeirra var breytt í annað vitlaust orð (25,5%)
  - 34 þeirra var ekki breytt (8,3%)

  - 20 þeirra voru ekki leiðrétta vegna þess þau virtust vera rétt (4,9%)
  - 7 þeirra voru ekki leiðrétt vegna þess að rétta orðið vantaði í listann (1,7%)

Eftir leiðréttingu voru rétt orð 1812/2013 (**90,01%**).

Í ÍBN 19s-20f eru 4,9 milljónir orða, 980 þúsund villur?

Stoðgögn, samanburðarefnið

- Beygingarlýsing íslensks nútímamáls (BÍN):
  u.þ.b. 5,8 milljónir orðmynda úr nútímamáli

- Textasafn Orðabókar Háskólans: tæplega 1,3 milljónir orðmynda, frá ýmsum tímum, með tíðnitölum.

- Ritmálssafn Orðabókar Háskólans:
  Orðmyndalisti úr dæmasafninu, ríflega 805 þúsund orðmyndir frá 16.-19. öld, með tíðnitölum.

Allar orðmyndir sem ekki koma fram í þessum gögnum eru kandídatar í villur!

Stoðgögnin eru á tilraunastigi enn sem komið er.

19. aldar stafsetning flækir málin verulega,
sjá sýnishorn úr lista úr Ritmálssafni →

Orðmyndir úr Ritmálssafni Orðabókar Háskólans

	16. öld	17. öld	18. öld	19.öld	BÍN	Textas.
keyfft	1	3			-	-
keyffte	2				-	-
keyffti	1				-	-
keyft	1			4	+*	1* (keyfa)
keyfti				3	-	1
keift			1		-	-
keifti				1	-	-
keijpti		1			-	-
keypte	3	8	3		-	3
keypter	2				-	3
keypt	9	31	32	328	+	5.077
keypti	8	17	14	213	+	4.265

Orðmyndir úr Ritmálssafni Orðabókar Háskólans

	16. öld	17. öld	18. öld	19.öld
amtamaður		1		
amtmadr				10
Amtmadrinn			2	
amtmaðr			1	23
Amtmaðr				2
Amtmaðrinn				1
amtmaðrinn				2
amtmaður		11		
Amtmaður		3	1	16
amtmaður			24	133
amtmaðurinn		5	18	38
Amtmaðurinn				9

Óviðráðanlegar villur

- **Prentvillur** sbr. sýnishorn áðan víð f. við
- **Umbrotsvandamál,** dálkabrengl o.þ.h.
- **Prentgæði, letur, brot eða bylgjur í pappír** o.þ.h.

Í handleiðréttu textunum er nákvæm villugreining.

Villur af þessu tagi getur við ekki leiðrétt vélrænt,
a.m.k. ekki á þessu stigi
Vandamál í afbrigðilegri stafsetningu eru líka illleysanleg.

Einföld aðferð til leiðréttingar

- Vinnum orðtíðnilista úr stóru textasafni
  - Öll orð sem ekki eru í tíðnilistanum teljast vera villur
- Leiðrétting:
  - Finnum öll orð í tíðnilistanum sem eru einni breytingu frá upphaflega orðinu. Breyting getur verið að:
    - Fjarlægja einn staf
    - Bæta við einum staf
    - Breyta einum staf í annan
    - (Víxla samliggjandi stöfum)
  - Skilum algengasta orðinu sem við fáum út.
  - Ef ekkert orð finnst endurtökum við leikinn, nú með tveimur breytingum í stað einnar.
  - Ef við finnum enn ekkert orð látum við upphaflega orðið standa eins og er.

Tvö dæmi um leiðréttingu

Dæmi: peir (þeir)

- Finnst ekki í orðtíðnilistanum
- Mögulegar leiðréttingar (ein breyting):
  - þeir 194.192
  - meir 5.060
  - geir 1.901
  - leir 873
  - ...
- Segjum því að **þeir** sé líklegasta leiðréttingin

Dæmi: útliinda (útlanda)

- Finnst ekki í orðtíðnilistanum
- Ekkert orð í tíðnilistanum er einni breytingu frá því
- Mögulegar leiðréttingar (tvær breytingar):
  - útlanda 821
  - útlenda 257
  - útlimina 18
  - útleidda 2
  - ...
- Segjum því að **útlanda** sé líklegasta leiðréttingin

Skilvirkni

Tímarit	Orð	Villur	%Rétt fyrir	Leiðrétt	Nýjar villur	%Rétt eftir	%Skilvirkni
Norðlingur (07.05.1881)	2072	442	78,67%	251	52	88,27%	56,79%
Þjóðólfur (29.01.1875)	2653	916	65,47%	441	64	79,68%	48,14%
Íslendingur (25.01.1875)	1831	242	86,78%	133	40	91,86%	54,96%
Tíminn (05.08.1922)	1953	172	91,19%	92	42	93,75%	53,49%
Vísir (23.12.1910)	1189	143	87,97%	54	40	89,15%	37,76%

Hvað er hægt að bæta?

- Við athugum ekki samhengið sem orðin eru í
  - Sum orð er einungis hægt að lagfæra með því að skoða orðin í kring.
    - Orðið **finn** lítur eitt og sér ekki út fyrir að vera rangt
    - Vitum þó að það sé vitlaust ef það kemur fyrir í setningunni “Maturinn var **finn** [...]”
- Stundum þarf að gera fleiri breytingar
  - Það gæti þurft að gera tvær breytingar til að fá rétta orðið, jafnvel þó við finnum einhverjar mögulegar leiðréttingar sem eru bara einni breytingu frá.
  - Sum orð eru fleiri en tveimur breytingum frá vitlausa orðinu.

- Allar breytingar eru jafnlíklegar
  - Algengt er að **þ** sé lesið inn sem **p** (t.d. verður **þeir** oft að **peir**). Þetta kemur ekki á óvart, enda eru stafirnir svipaðir í útliti og það getur verið erfitt að greina á milli þeirra.
  - Þeir var líklegasta leiðréttingin á peir, en önnur orð komu líka til greina, t.d. geir, meir og leir. Það er þó fátt líkt með **p** og **g**, **m** eða **l**. Það hljóta því að vera mun ólíklegri breytingar.

Samanburður á villutíðni

Dagskrá (13.07.1896)

Stafur	Fjöldi	Villur	%Rétt
a	1554	9	99,42%
ð	745	6	99,19%
l	747	27	96,39%
ó	139	22	84,17%
s	884	41	95,36%
t	833	1	99,88%
þ	231	0	100,00%
æ	118	0	100,00%
ö	133	6	95,49%

Íslendingur (25.01.1875)

Stafur	Fjöldi	Villur	%Rétt
a	2380	1	99,96%
ð	1222	5	99,59%
l	1069	15	98,60%
ó	264	3	98,86%
s	1282	1	99,92%
t	1186	11	99,07%
þ	446	93	79,15%
æ	193	2	98,96%
ö	206	3	98,54%

Líkindi á breytingum

- Það fer eftir ástandi og eiginleikum upprunalega textans hversu líklegar tilteknar breytingar eru.
  - Getum t.d. ekki gefið okkur fyrirfram að það séu einhverjar fastar líkur á því að **p** ætti að vera **þ**.
- Hvernig eigum við þá að vita hvaða vægi við ætlum að gefa breytingunum?
  - Ef við getum leiðrétt um 70% af villunum, fáum við ekki nokkuð góða mynd af því hvaða villur hafa verið gerðar?

Hugmynd

1. Leiðrétta textann og safna jafnóðum saman upplýsingum um hvaða breytingar voru gerðar.
2. Leiðréttum textann aftur, en notum nú tíðnina á villunum til að gefa mismunandi breytingum vægi eftir því hversu algengar þær eru. Höldum áfram að safna upplýsingum um tíðni breytinga.
3. O.s.frv.

Villa	Leiðrétting	Breyting
pessir	þessir	p -> þ
Reykjavik	Reykjavík	i -> í
íleiri	fleiri	í -> f
stgrkja	styrkja	g -> y

c(peir)	Tíðni	Breyting
þeir	194.192	p -> þ (32)
meir	5.060	p -> m (0)
geir	1.901	p -> g (0)
leir	873	p -> l (0)

Skilvirkni

Tímarit	%Rétt	Leiðrétting 1	Leiðrétting 2	%Eff 1	%Eff 2
Norðlingur 07.05.1881	78,67%	88,27%	89,33%	56,79%	61,76%
Þjóðólfur 29.01.1875	65,47%	79,68%	80,55%	48,14%	50,66%
Íslendingur 25.01.1875	86,78%	91,86%	92,68%	54,96%	61,16%
Tíminn 05.08.1922	91,19%	93,75%	93,80%	53,49%	54,07%
Vísir 23.12.1910	87,97%	89,15%	89,32%	37,76%	39,16%

Framhaldið

- Textarnir 625 í ÍBN 19s-20f verða handleiðréttir í áföngum
- Þessir leiðréttu textar verða notaðir til vélrænnar leiðréttinga
- Stoðgögnin þarf að bæta
    - Bæta þarf við meiri gögnum!
    - Tíðnitölur þarf að vinna betur, t.d. athuga hvað orðmynd þarf að vera algeng til að vera nothæf
    - Skoða þarf stafsetningu tímabilsins sérstaklega, m.t.t. stoðgagnanna
- Þessir leiðréttu textar verða notaðir til vélrænnar leiðréttingar

- **Og leiðréttingarbúnaðurinn verður áfram í þróun!**

Takk fyrir áheyrnina!

Jón Friðrik Daðason **jfd1@hi.is**
Kristín Bjarnadóttir **kristinb@hi.is**

STOFNUN ÁRNA MAGNÚSSONAR
Í ÍSLENSKUM FRÆÐUM