Sigrún Helgadóttir
Stofnun Árna Magnússonar í íslenskum fræðum

Málheildir með íslenskum textum

Námskeiðið “Íslensk textaföng”
6. október 2014

Yfirlit

Hvað er málheild? Skilgreining
Mismunandi gerðir málheilda
Til hvers má nota málheildir?
Erlendar málheildir
Líkindalíkan, tölfræðilegt mállíkan, n-stæður
Listi yfir helstu málheildir (textasöfn) og stutt lýsing á þeim. Annað efni í málheildir
Nánar um Íslenska orðtíðnibók, Íslenskan orðasjóð, MerkOr, Gullstaðalinn, MÍM
Málheild með eldri textum
Leitarkerfi MÍM, dæmi
Framtíðarmúsík
Notkun málheilda í verkefnum – gestir

Sagnfræði

Upphaf máltækni á Íslandi

Skýrsla frá 1999 fyrir menntamálaráðherra (RÓ, ER, ÞS)

Tungutækniverkefni menntamálaráðuneytisins 2000–2004

BÍN, mörkun, MÍM, talgervill, talgreining, bættur Púki, menntun,...

Verkefnið fékk um 130 M en hætti svo.

Verkefni háð fjárveitingum stofnana og styrkjum frá Rannís, Rannsóknasjóði HÍ, Nýsköpunarsjóði námsmanna, Þjóðhátíðarsjóði, ….

Nokkrir fræðimenn sóttu námskeið í Gautaborg og fengu þar hugmyndir – mörkunarverkefnið afsprengi námskeiðsverkefnis, markaðir textar

Orðtíðnibókarinnar fundust í tölvukerfi OH (þjálfun µ-TBL, transformation based learning)

Hvað er málheild (e. corpus)?

Safn stafrænna texta sem hafa tiltekið snið

-Textasafnið þarf að vera nógu stórt til að vera á einhvern hátt lýsandi fyrir tungumálið

-Getur verið hreinn texti

-Geta verið markaðir (beygingarfræðileg og/eða setningafræðileg mörk) og/ eða lemmaðir

-Geta fylgt þeim lýsigögn (e. *metadata*)

-Geta verið marglaga – hliðstæðir textar á milli tungumála eða málstiga

Snið textanna getur verið misjafnt á milli textasafna en innan sama textasafnsins eru textarnir í sama sniði.

Málheildir eru oft geymdar í xml-sniði

(Byggt á: Steinþór Steingrímsson: Atrenna að textasöfnum: Úr smiðju og hugskoti Árnastofnunar, Hugvísindaþingi 15. mars 2014)

Textaval fyrir málheildir

Málheildir skiptast í tvo flokka eftir textavali:

Sérhæfð málheild

- Dæmi: Dagblaðatextar, textar um læknisfræði o.s.frv.
- Notkun: kanna málfar á tilteknu sviði; búa til sérhæfð máltæknitól, t.d. talgreini fyrir röntgenlækna

Málheild með fjölbreyttum textum (*balanced*)

- Dæmi: Mörkuð íslensk málheild (MÍM)
- Notkun: kanna almennt málfar; búa til máltæknitól fyrir almenna notkun

Tímarammi texta

Málheildir hafa efni frá einu tímabili (*synchronous*) eða frá ýmsum tímum (*diachronic*)

Samtíma málheild (*synchronous*)

Allir textar skrifaðir á tilteknu tímabili, t.d. 2000-2009 (*MÍM*)

Notkun: kanna málfar á tilteknu tímabili, búa til máltæknitól fyrir tiltekið tímabil

Söguleg málheild (*diachronic*)

Textar frá ýmsum tímum. Dæmi: Íslenski trjábankinn (*IcePaHC*)

Notkun: kanna málbreytingar í tíma

Hvernig eru málheildir búnar til?

Skref við undirbúning málheildar:

1. Texti er valinn í samræmi við hvers konar málheild á að búa til
2. Afla leyfa frá rétthöfum ef um það er að ræða, velja eða búa til leyfir fyrir dreifingu textanna
3. Textinn er dreginn úr sniðinu sem hann er í (umbrotssnið, pdf, xml,…)
4. Textinn er hreinsaður (efnisyfirlit, neðanmálsgreinar, “vond” tákn, o.s.frv.)
5. Ákveða hvaða lýsigögn á að skrá með hverjum textabút og hvernig orð í textanum eru merkt, og snið textanna í málheildinni
6. Greina texta í setningar og lesmálsorð
7. Marka texta og finna nefnimyndir, frekari greining,…
8. Búa textana til dreifingar: textar settir t.d. í xml-snið (með lýsigögnum og mörkum) fyrir þá sem vilja fá textana til rannsókna og fyrir gerð máltæknitóla og gert leitarviðmót

Til hvers eru málheildir notaðar?

Með þeim er hægt að gera tölfræðilega greiningu á tungumálinu
-Tíðni orða, orðmynda, setningargerða o.s.frv. sem geta nýst við orðabókargerð eða gerð kennsluefnis

-Til að prófa fræðilegar tilgátur. T.d. um reglur í málfræði eða setningarfræðileg fyrirbæri.

-Til að búa til líkindalíkan fyrir talgreiningu, sjálfvirkar þýðingar, mörkunarforrit, leiðréttingartól og fleiri máltæknitól (flokkun texta, greina tungumál o.s.frv., sjá frekari dæmi frá JFD)

Undanfarið hefur færst í aukana að nota textasöfn í félagsmálfræði t.d. til þess að skoða breytileika í máli milli staða og mismunandi þjóðfélagshópa.

Erlendar málheildir

Elstu málheildirnar hafa enska texta. Sennilega er elsta málheildin **Brown Corpus** frá um 1961, hafði ríflega eina milljón orða af margvíslegum textum, hver texti hafði um 2000 orð (fyrirmynd Íslenskrar orðtíðnibókar).
BNC (British National Corpus): 1991–1994, 100 M orð, 10% talmál
Hefur verið fyrirmynd flestra málheilda sem gerðar hafa verið síðar (t.d. MÍM)
ANC (American National Corpus) og OANC (Open American National Corpus): ANC 22 milljón orð, OANC um 15 M orð, opnari. Málfræðileg mörk, nefnimyndir, setningagreining, nafnaþekking (mikil greining)
COAC (Corpus of Contemporary American English): 450 M orð, 160 þús. textar frá 1990–2011. Leitarkerfi BNC og mörkunarforrit (CLAWS)

Málheildir hafa verið gerðar fyrir mörg tungumál í evrópu og annars staðar, sjá t.d. **National Corpus of Polish**: á að hafa einn milljarð orða, hefur nú um 300 milljónir. Til eru málheildir fyrir öll norrænu málin, misjafnlega aðgengilegar.

N-stæður

n-stæða (n-gram) er samfelld runa n atriða úr texta eða tali. Atriðin geta verið hljóðön (e. *phonems*), atkvæði, orðmyndir, nefnimyndir (e. *lemmas*), stafir, mörk,…

Tíðni n-stæðna er fundin til þess að búa til tölfræðileg mállíkön (líkindalíkön) fyrir tiltekin tungumál.

n=1 → „unigrams“; einstæður
n=2 → „bigrams“; tvístæður
n=3 → „trigrams“; þrístæður

Eftir því sem n er hærra þarf stærri málheild þar sem ýmsar langar runur koma ekki fyrir í minni málheildum („sparse data problem)

Líkindalíkan – tölfræðilegt mállíkan

Tölfræðilegt mállíkan er tölfræðilegt líkan sem spáir fyrir um næsta orð á grundvelli n-1 orða sem fara næst á undan
Líkindin $P(w_1,\ldots,w_n)$ á því að sjá setninguna $w_1,\ldots,w_n$ eru nálguð með

$$P(w_1,\ldots,w_m) = \prod_{i=1}^{m} P(w_i \mid w_{i-(n-1)},\ldots,w_{i-1}) \approx \prod_{i=1}^{m} P(w_i \mid w_{i-(n-1)},\ldots,w_{i-1})$$

Skilyrtu líkurnar má reikna út frá tíðni n-stæðna

$$P(w_i \mid w_{i-(n-1)},\ldots,w_{i-1}) = \frac{count(w_{i-(n-1)},\ldots,w_{i-1},w_i)}{count(w_{i-(n-1)},\ldots,w_{i-1})}$$

Sjá t.d. Google N-Grams Corpus
http://storage.googleapis.com/books/ngrams/books/datasetsv2.html
og N-stæðu verkefni Kristjáns Rúnarssonar

Aðgengileg textasöfn á málföng.is

Íslensk orðtíðnibók

-Rúmlega 590 þúsund lesmálsorð greind málfræðilega, mörk og nefnimyndir leiðrétt

-Leitaraðgangur og u.þ.b. 70% texta aðgengilegir til nota í rannsóknum og máltækniverkefnum (engir þýddir textar), allir textar aðgengilegir fyrir þjálfun markara (fyrir 10-fold cross-valdidation)

Mörkuð íslensk málheild (MÍM)

-25 milljón lesmálsorð greind málfræðilega, mörk og nefnimyndir ekki leiðrétt

-Leitaraðgangur og allir textar aðgengilegir í xml-sniði til nota í rannsóknum og máltækniverkefnum. Tíðni nefnimynda sem koma fyrir oftar en 100 sinnum.

Aðgengileg textasöfn á málföng.is (frh.)

Gullstaðallinn (MIM-GOLD)

-Um einnar milljónar orða úrtak úr textum MÍM, mörkun hefur verið leiðrétt. Útgáfa 0,9 er aðgengileg. Nú útgáfa kemur í vetur.

Fornritatextar

-Um 1,6 milljónir orða af textum úr fornritum, textar markaðir. Aðgengileg til leitar og til nota í xml-sniði fyrir rannsóknir og máltækniverkefni.

Íslenskur trjábanki (Icelandic Parsed Historical Corpus – IcePaHC)

Um 1 milljón orða af setningagreindum textum frá 12. öld til 21. aldar.

-Alveg opinn. Sjá http://www.linguist.is/icelandic_treebank/Download

Aðgengileg textasöfn á málföng.is (frh.)

Íslenskur orðasjóður

Unninn í háskólanum í Leipzig. Um 550 milljónir lesmálsorða, fyrst og fremst veftextar.

MerkOr - Íslenskur merkingarbrunnur

Íslenskt orðasafn sem byggist á merkingarvenslum milli orða og flokkun þeirra í merkingarsvið. Allt innihald MerkOr varð til með sjálfvirkum aðferðum.

Hugtakasafn Þýðingarmiðstöðvar utanríkisráðuneytisins

Íðorð sem hafa verið notuð sl 25 ár í þýðingum á lagatextum sem falla undir samninginn um Evrópska efnahagssvæðið (EES-samninginn). Nú eru í safninu 70.000 færslur. Sjá http://www.hugtakasafn.utn.stjr.is/

Önnur textasöfn Árnastofnunar

Íslenskt textasafn

Um 65 milljónir orða af gömlum og nýjum textum af ýmsum toga. Opið fyrir leitaraðgang. Sjá http://corpus.arnastofnun.is/leit.pl?info=1 og fyrirlestur Þórdísar Úlfarsdóttur 15. sept. sl.

Ritmálssafn

Dæmi um notkun orða frá miðri 16. öld og fram til loka 20. aldar

2,5 milljónir notkunardæma, 707 þúsund uppflettiorð, rúmlega 20 milljónir uppflettiorða (efni í „Orðabókina“)

Sjá http://www.arnastofnun.is/page/gagnasofn_ritmalssafn

Hugsanlegt efni í málheildir

Þjóðháttatextar Þjóðminjasafnsins

Frá sjöunda áratug 20. aldar og fram yfir aldamót, 16 milljónir orða. Erfitt að fá leyfi fyrir notkun

Ræður Alþingismanna

Um 190 milljónir orða af ræðum frá 1950. Auðvelt að sækja og þarf ekki leyfi til þess að nota.

Textar úr gagnasafni Morgunblaðsins

Um 300 milljónir orða af textum úr Morgunblaðinu frá 1987. Þarf leyfi en tiltölulega auðvelt að afla þess fyrir rannsóknarverkefni. Í MÍM eru um 5 milljónir orða úr þessu safni.

Hugsanlegt efni í málheildir

Wikipedia

Íslenskur hluti Wikipediu, 5 milljónir orða (mars 2014). Opið og frjálst til notkunar

Biblíutextar

-11 biblíuþýðingar á íslensku frá mismunandi tímum

-E.t.v. 2-3 aðgengilegar stafrænar

Opið og frjálst?

Útgáfuréttur á nýjustu þýðingunni

Hjá Háskólanum í Edinborg (http://homepages.inf.ed.ac.uk/s0787820/bible/) eru aðengilegar í xml-sniði þýðingar á ótalmörg tungumál, m.a. Íslensku

Sú gerð er tekin af bandarískri síðu https://www.biblegateway.com/versions/ og er sennilega þýðingin frá 1981

Flestar (e.t.v. allar) biblíuþýðingar á http://baekur.is/.

Mætti nota til þess að skoða breytingar á máli á 400 ára tímabili, en mikil vinna að undirbúa textana

Önnur textasöfn

Landsbókasafnið

-Timarit.is (sjá 19. aldar verkefni Árnastofnunar)

-Skemman

-Rafhlaðan

-Bækur.is

Veftextar

-Bland.is

-Fréttir og blogg

-Athugasemdir við fréttir og blogg

-Facebook og Twitter

-..

Íslensk orðtíðnibók

Íslensk ortíðnibók, gefin út 1991 af Orðabók háskólans, ritstjórar Jörgen Pind, Friðrik Magnússon og Stefán Briem

Búin til sérstök málheild til þess að finna tíðnitölur

Fyrirmynd er “Brown corpus” og lík verkefni

Málheild Orðtíðnibókarinnar hefur verið notuð sem gullstaðall fyrir þjálfun og prófun námfúsra markara.

Íslensk orðtíðnibók – málheildin

Val á textum í málheild Íslenskrar orðtíðnibókar

1. 100 textabrot, hvert um 5.000 lesmálsorð
2. Valið úr ritverkum sem voru fyrst gefin út 1980–1989.
3. Fimm textaflokkar:

(a) Íslensk skáldverk
(b) Þýdd skáldverk
(c) Ævisögur og endurminningar
(d) Fræðslutextar (skipt jafnt milli fræðsluefnis á sviði raunvísinda og tækni (10 textar) og fræðsluefnis á sviði hugvísinda (10 textar))
(e) Barna- og unglingabækur (frumsamið íslenskt efni (10 textar) og þýtt efni (10 textar)

4. Höfundur eða þýðandi hvers texta má ekki vera höfundur eða þýðandi annars texta

Íslensk orðtíðnibók – málheildin

Val á textum í Íslenska orðtíðnibók

- Hver texti hefst á samfelldu máli
- Texti er venjulega tekinn úr upphafi meginmáls, myndefni og öðru sundurlausu efni sleppt
- Fyrirsögnum og kaflanúmerum inni í texta er ekki sleppt
- Texti endar á heilli setningu, venjulega við lok kafla eða annars skýrt afmarkaðs rithluta
- Val texta réðst yfirleitt af því hversu auðfenginn textinn var. Flestir textarnir voru tiltækir í textasafni Orðabókar, nokkra þurfti að skrá sérstaklega
- Leyfa ekki aflað frá rétthöfum (gert í tengslum við MÍM síðar)

Íslensk orðtíðnibók – málheildin

Mörkun

I: 54.000 lesmálsorð voru handmörkuð í tilraunaverkefni (sjá Friðrik Magnússon. 1988. „Hvað er títt?“ *Orð & tunga* 1:1–49.)

II: Sjálfvirk mörkun sem byggðist á handmörkun í I; 50 textar, 250.000 lesmálsorð. Leiðrétt í höndum.

III: Sjálfvirk mörkun með því að nota endurbætur á markaranum sem byggðust á II; 50 textar, 250.000 lesmálsorð

„Bootstrapping“

Íslensk orðtíðnibók – málheildin

Mörkun

- Stefán Briem samdi markarann, byggðist á beygingarfræðilegum atriðum, setningafræðilegum og tölfræðilegum
- Áætluð nákvæmni um 80% (gaf bæði greiningarstreng og nefnimynd)
- Um 1990 var markarinn endurbættur og náði þá um 90% nákvæmni

Íslensk orðtíðnibók – málheildin

Markaskrá (sjá ljósrit)

- Byggir á hefðbundinni greiningu á íslensku í orðflokka og málfræðilega flokka
- Sums staðar hefur hefðbundinni greiningu verið breytt vegna setningafræðilegrar stöðu orða
- Forsetningar er t.d. greindar sem atviksorð

Íslensk orðtíðnibók – mörkin

orð	nefnimynd	mark	skýring
ég	ég	fp1en	f: fn; p: pfn; 1: 1. pers.; e: et.; n: nefnifall
stökk	stökkva	sfg1eþ	s: so.; f: frsh.; g: germ.; 1: 1. pers.; e: et.; þ: þátíð
á	á	aa	a: ao.; a: stýrir ekki falli
eftir	eftir	aþ	a: ao.; þ: stýrir þágufalli
strætó	strætó	nkeþ	n: no.; k: kk.; e: et.; þ: þgf.
og	og	c	c: samtenging
veifaði	veifa	sfg1eþ	s: so.; f: frsh.; g: germ.; 1: 1. pers.; e: et.; þ: þátíð
,	,	,	komma
vagnstjórinn	vagnstjóri	nkeng	n: no.; k: kk.; e: et.; n: nf.; g: með greini
sá	sjá	sfg3eþ	s: so.; f: frsh.; g: germ.; 3: 3. pers.; e: et.; þ: þátíð
mig	ég	fp1eo	f: fn.; p: pfn.; 1: 1. pers.; e: et.; n: þolfall
og	og	c	c: samtenging
stoppaði	stoppa	sfg3eþ	s: so.; f: frsh.; g: germ.; 3: 3. pers.; e: et.; þ: þátíð
.	.	.	punktur

Mynd 1. Greining orða í einni setningu úr skáldsögunni *Mín káta angist* eftir Guðmund Andra Thorsson

Íslensk orðtíðnibók – notkun

Textar Orðtíðnibókarinnar eru aðgengilegir á þrennan hátt:

1. Til leitar á http://mim.arnastofnun.is/. Heppilegt til þess að kenna orðflokkagreiningu þar sem greiningin er handleiðrétt
2. Sækja má textana í xml-sniði (þá sem leyfi hefur fengist fyrir) og nota í rannsóknum og máltækniverkefnum
3. Sækja má 10 pör af þjálfunar- og prófunarsöfnum og nota t.d. við þjálfun námfúsra (data-driven) markara

Íslensk orðtíðnibók – notkun

Málheild Orðtíðnibókarinnar hefur verið notuð til þess að þjálfa og þróa alla markara sem eru notaðir fyrir mörkun á íslensku máli.

Gallar:

Málheildin er heldur lítil – sum fyrirbæri koma ekki fyrir
Textar frá afmörkuðu tímabili (1980–1990)
Of mikil bókmenntaleg slagsíða, 80%

Grundvöllur að mörkun á íslensku þó að það hafi ekki verið markmiðið með gerð málheildarinnar

Íslensk orðtíðnibók – verkefni?

Búa til kennsluefni sem kennir orðflokkagreiningu?

Fleira?

Íslenskur orðasjóður - textarnir

Málheild með textum úr íslensku nútímamáli, fimm mismunandi heimildir, textar úr hverri heimild mynda undirmálheild

- WWW 2005: Vefsíðusöfnunum Landsbókasafns Íslands - Háskólabókasafns haustið 2005 (u.þ.b. 227 milljón lesmálsorð).
- WWW 2010: vefsíðusöfnunum Landsbókasafns Íslands - Háskólabókasafns haustið 2010 (u.þ.b. 366 milljón lesmálsorð).
- Dagblað 2002: Morgunblaðið 2001 (u.þ.b. 18,1 milljón lesmálsorð).
- Dagblað 2011: Textar úr netútgáfum íslensku dagblaðanna 2011 (u.þ.b. 22,6 milljón lesmálsorð).
- Wikipedia: textar úr íslenska Wikipedia-alfræðiritinu 2010 (u.þ.b. 2,5 milljón lesmálsorð).

Íslenskur orðasjóður - textarnir

Í málheildinni eru samtals:

- tæplega 33 milljón setningar
- 545 milljónir lesmálsorða (eftir að eins setningar hafa verið fjarlægðar),
- þar af eru 6,7 milljónir mismunandi orðmynda

Textarnir voru markaðir hjá Stofnun Árna Magnússonar haustið 2012 með sama kerfi og MÍM.
Mörkunin er ekki aðgengileg sem hluti af málheildinni enn þá.
Með því að hafa samband við höfunda má fá aðgang að textunum til notkunar í málrannsóknum og máltækniverkefnum.
Höfundar hafa gefið út tíðniorðabók.
http://wortschatz.uni-leipzig.de/ws_isl/

Íslenskur orðasjóður - leit

Leit á vefsíðunni http://wortschatz.uni-leipzig.de/ws_isl/
Leita má að orðmynd, engar nefnimyndir eru í málheildinni
Þetta er sýnt:

- Heildartíðni orðmyndar í textagrunninum.
- Tíðniflokk orðmyndar, þ.e. hlutfallslega tíðni orðmyndar miðað við og sem algengasta orð.
- Notkunardæmi og tengill við fleiri notkunardæmi.
- Lista með orðum sem hafa háa tíðni sem nágrannar leitarorðsins, þ.e. orð sem koma oft fyrir í sömu setningum og leitarorðið.
- Lista með orðum sem hafa háa tíðni sem vinstri nágrannar leitarorðsins, þ.e. orð sem koma oft fyrir sem næsta orð framan við leitarorðið innan sömu setningar.
- Lista með orðum sem hafa háa tíðni sem hægri nágrannar leitarorðsins, þ.e. orð sem koma oft fyrir sem næsta orð aftan við leitarorðið innan sömu setningar.
- Merkingarnet sem sýnir tengsl leitarorðsins við orð sem hafa marktæka tíðni í sömu setningum og leitarorðið.

Íslenskur orðasjóður - notkun

- Pólska fyrirtækið Ivona sem bjó til talgervilinn Karl/Dóru notaði texta Orðasjóðsins
- Jón Friðrik Daðason hefur notað textana við gerð Skramba (sjá síðar)
- Höfundar hafa gert tíðnirannsóknir á mörkuðum textum til samanburðar við niðurstöður í Íslenskri orðtíðnibók (sjá **A 500 Million Word POS-Tagged Icelandic Corpus**)
- Verkefni: ?????

MerkOr - íslenskur merkingarbrunnur

MerkOr er nýstárlegt íslenskt orðasafn sem byggist á merkingarvenslum milli orða og flokkun þeirra í merkingarsvið. Allt innihald MerkOr varð til með sjálfvirkum aðferðum sem eiga að nýtast til þess að búa til fleiri orðasöfn af þessu tagi, jafnt fyrir íslensku sem önnur tungumál. Meðal heimilda við gerð gagnagrunnsins voru skýringar úr Íslenskri orðabók.

Notkun

Leita má í gögnunum
(http://merkor.skerpa.com/MerkorApplication)
og
Sækja þau (https://github.com/bnika/MerkOrCore)
Sjá nánar á http://málföng.is/
Höfundur: Anna Björk Nikulásdóttir

MerkOr - íslenskur merkingarbrunnur

Notkun ????

Verkefni????

Gullstaðallinn – MIM-GOLD

Úrtak úr MÍM með um einni milljón lesmálsorða. Mun taka við eða verða til viðbótar við málheild Íslenskrar orðtíðnibókar.
Texti úr 13 textaflokkum:

Textaflokkur	fj. orða	%
Morgunblaðið	248.879	24,73
Bækur	237.065	23,56
Blogg	135.489	13,46
Fréttablaðið	94.487	9,39
www.visindavefur.is	92.202	9,16
Vefsetur	65.177	6,48
Lög	41.217	4,10
Skólaritgerðir	34.357	3,41
Til upplestrar	19.354	1,92
Dómar	12.936	1,29
Fréttir útvarps og sjónvarps	11.194	1,11
Vefmiðlar	8.524	0,85
Tölvupóstur	5.512	0,55
Samtals	1.006.393	100,00

Gullstaðallinn – MIM-GOLD

•Þróað var kerfið *WorkerBranch (*nú kallað *CorpusTagger)* fyrir tilreiðslu, mörkun og lemmun, þetta kerfi hefur síðan verið notað fyrir *MÍM* og *Orðasjóðinn*.

•Í upphaflega kerfinu var notaður tilreiðari úr *IceNLP.* Markað var með fimm mörkurum (*IceTagger*, *TnT*, *MXPOST*, *fnTBL* og *Bidir*) og síðan kosið á milli markanna með forritinu *CombiTagger*.

•*TnT*, *MXPOST*, *fnTBL* eru námfúsir markarar sem voru þjálfaðir á málheild Íslenskrar orðtíðnibókar en IceTagger er reglumarkari sem var einnig gerður með því að nota markaða texta Orðtíðnibókarinnar. Við gerð Bidir var líka stuðst við Orðtíðnibókina. Orðalisti úr BÍN er notaður fyrir suma markarana til þess að bæta mörkun óþekktra orða.

•Forritið *Lemmald* er notað til þess að finna nefnimyndir

•Bidir markarinn var fjarlægður úr kerfinu þar sem hann réði ekki við stórar skrár og TriTagger (Hrafn Loftsson) kom í staðinn fyrir TnT

Gullstaðallinn – leiðréttingaferli

1. Leitað var að villum með því að grunnþátta textana með IceParser og athuga ósamræmi í nafnliðum, sagnliðum og forsetningarliðum
2. Villurnar voru leiðréttar og nákvæmni mörkunar var metin með því að skoða mark um hundraðasta hvers orðs. Mörkunarnákæmni var metin 88–95%.
3. Síðan var farið var yfir hvert mark og mörk leiðrétt (Kristján Friðbjörn Sigurðsson). Nákvæmni mörkunar metin um 91,5-98,5% með því að skoða hundraðasta hvert orð.
4. Orð í málheildinni voru mörkuð með IceTagger. Þau mörk voru borin saman við “rétt” mörk, ef mismunur fannst voru þau orð merkt (Steinunn Valbjörnsdóttir o. fl.)

Gullstaðallinn – leiðréttingaferli

5. Eitt af þrennu var valið

(i) “rétt” mark í málheildinni

(ii) mark IceTagger

(iii) nýtt rétt mark

Nákvæmni mörkunar eftir þennan verkþátt var metin um 98,5-100,0% með því að skoða hundraðasta hvert orð.

Leiðréttar skrár eftir fyrstu handleiðréttingu má sækja á http://málföng.is/ (3 að ofan, útgáfa 0,9)

Stefnt er að því að veita aðgang að skránum eftir seinni leiðréttingu á næstu mánuðum. Gerð verða 10 pör fyrir þjálfun og prófun eins og gert var fyrir skrár Orðtíðnibókarinnar.

Gullstaðallinn – meira

Eitt af því sem þarf að gera þegar skrár Gullstaðalsins eru tilbúnar markaðar er að finna nefnimyndir. Forritið *Lemmald* er frekar ónákvæmt. Nýtt lemmunarforrit, Nefnir, er í smíðum (Jón Friðrik Daðason). Þegar það verður tilbúið verða fundnar nefnimyndir fyrir orð í textum Gullstaðalsins. Óhjákvæmilega verða þar villur. Nauðsynlegt verður að leiðrétta þær villur eins og kostur er.

Mætti nota hópvirkjun (“crowdsurcing”)?

Væru nemendur t.d. tilbúnir til þess taka þátt í þannig verkefni?

Gullstaðallinn – meira

Notkun:

Gullstaðallinn verður notaður til þess að þjálfa námfúsa markara, bæði þá sem þegar hafa verið þjálfaðir og einnig markara sem hafa orðið til síðan mörkunarverkefnið var unnið.

Verkefni:

Finna fleiri leiðir til þess að finna villur í mörkun og bæta þannig nákvæmni

Gera tíðnikönnun, sambærilega við það sem er í Orðtíðnibókinni

Mörkuð íslensk málheild (MÍM)

Hvers konar málheild?

Skyldi hafa um 25 milljónir orða af fjölbreyttum (balanced) textum frá 21. öld (synchronous). Hluti af textunum skyldi vera talmál. Hverjum texta skyldi fylgja lýsigögn (metadata) og hverju orði málfræðilegar upplýsingar og nefnimynd.

Leyfa skyldi aflað frá rétthöfum texta sem voru varðir af höfundarrétti til þess að leyfa sem víðasta notkun.

Aðferðir við gerð BNC voru hafðar til viðmiðunar. Notkunarleyfið er sniðið eftir notkunarleyfi BNC.

Nú er algengast að nota stöðluð notkunarleyfi (t.d. Creative Commons leyfi, https://creativecommons.org/))

Mörkuð íslensk málheild (MÍM)

- Textaöflun tók mið af því hvaða textar voru aðgengilegir. Safnað var ríflega 25 milljónum lesmálsorða af frumsömdum textum frá árunum 2000–2009 eftir höfunda sem hafa íslensku að móðurmáli úr 23 textaflokkum. 2% af textunum er talmál.
- Aðeins var safnað textum sem voru aðgengilegir í rafrænu formi.
- Leyfis var aflað hjá rétthöfum til þess að fá að nota alla texta í safninu sem voru varðir af höfundarrétti. Samið var um að textar úr útgefnum bókum væru styttir um 20%; aðrir textar eru óstyttir.

MÍM – höfundarréttur

- Leyfis var aflað til þess að fá að nota alla texta (að undanskildum textum frá opinberum aðilum sem eru ekki verndaðir af höfundarrétti, 9. gr. 73/1972)
- Rétthafar fengu upplýsingabækling um málheildina, uppkast að notkunarleyfi og samþykkisyfirlýsingu
- Rétthafar prentaðra bóka fengu gögn í pósti, aðrir um tölvupóst
- Ein áminning var send
- Gert var samkomulag við Rithöfundasamband Íslands, Hagþenki og Félag íslenskra bókaútgefenda um gerð málheildlarinnar

MÍM – höfundarréttur

Helstu atriði notkunarleyfis:

Þeir sem fá texta MÍM til notkunar í rannsóknum eða fyrir máltækniverkefni mega ekki framselja textana til þriðja aðila. Leyfishafi má ekki gefa textana út eða veita aðgang að þeim á neinn hátt eða hafa af þeim tekjur.

Leyfishafinn má nota það sem hann lærir af málheildinni á hvern þann hátt sem hann kýs. Má þar t.d. nefna að gera má n-stæður úr textum málheildarinnar og nota við gerð máltæknitóla (t.d. til þess að spá fyrir um hvert næsta orð er í innslegnum texta - “word prediction”)

Textaflokkar í Markaðri íslenskri málheild (MÍM)	Fjöldi orða	%
Textar úr prentuðum bókum	5.972.893	23,89
Morgunblaðið	5.019.617	20,08
Prentuð tímarit af ýmsu tagi	2.379.848	9,52
Blogg	1.976.706	7,91
Pistlar af Vísindavef	1.838.909	7,36
Skýrslur og greinargerðir af vefsetrum ráðuneyta[1]	1.695.304	6,78
Texti af vefsetrum fyrirtækja, samtaka og stofnana	1.337.764	5,35
Dómar[1]	886.240	3,54
Ræður fluttar á Alþingi[1]	526.444	2,11
Fréttablaðið	514.189	2,06
Talmál	504.318	2,02
Lokaritgerðir háskólastúdenta	486.677	1,95
Efni til upplestrar	432.287	1,73
Frumvörp og lög af vef Alþingis[1]	406.002	1,62
Fréttir útvarps og sjónvarps (RÚV)	262.219	1,05
Vefmiðlar	245.703	0,98
Stúdentsprófsritgerðir í íslensku	179.365	0,72
Veftímarit	121.374	0,49
Tölvupóstlistar	120.312	0,48
Textavarp	45.887	0,18
Texti úr tónleikaskrám Sinfoníuhljómsveitar Íslands	24.832	0,10
Kvikmyndadómar úr Morgunblaðinu	15.682	0,06
Safnaðarblöð	7.950	0,03
Samtals	**25.000.522**	**100,00**
[1]Textar sem eru ekki varðir af höfundarrétti	3.513.990	14,06
Textar varðir af höfundarrétti	21.486.532	85,94

MÍM – textar

Textar úr prentuðum bókum: leyfi frá höfundum, textar frá útgefendum eða höfundum, lýsigögn skráð

Morgunblaðið: textar fengnir úr gagnasafni Morgunblaðsins, lýsigögn fylgdu

Prentuð tímarit af ýmsu tagi: texti frá útgefendum (Heimur, Birtingur,…), af vef, frá höfundum,…lýsigögn skráð

Blogg: safnað af bloggsíðum með leyfi höfunda (stjórnmálamenn, guðfræðingar, almennir bloggarar)

Pistlar af Vísindavef: fengust frá ritstjórn http://visindavefur.is/, ritstjóri aflað leyfis höfunda

Skýrslur og greinargerðir af vefsetrum ráðuneyta: tekið af vefjum 14 ráðuneyta, ekki þurfti leyfi

Texti af vefsetrum fyrirtækja, samtaka og stofnana: af vefjum 14 fyrirtækja og stofnana, leyfi frá forstöðumönnum

MÍM – textar

Dómar: af vefjum Hæstaréttar og Héraðsdóms Reykjavíkur, ekki þurfti leyfi

Ræður fluttar á Alþingi: af vef Alþingis (úr gagnasafni þingsins með aðstoð tölvumanns Alþingis), ekki þurfti leyfi

Fréttablaðið: pdf-skjöl af vef Fréttablaðsins með leyfi ritstjórnar (ekki tókst að fá efni úr gagnasafninu)

Talmál: Umritaðir textar fengnir úr fjórum sjálfstæðum rannsóknarverkefnum, texti og tal Alþingisræðna verður gert aðgengilegt á leitarsíðu MÍM

Lokaritgerðir háskólastúdenta: fengnar ásamt leyfi frá höfundum

Efni til upplestrar: útvarpserindi (Íslenskt mál og fl.), predikanir og hugvekjur presta, ávörp, fengið með leyfi frá höfundum

Frumvörp og lög af vef Alþingis: tekið af vefnum, ekki þurfti leyfi

Fréttir útvarps og sjónvarps (RÚV): fengið fyrir tilstyrk starfsmanna RÚV af vef Miðlunar (náðist ekki úr gagnasafni RÚV)

MÍM – textar

Vefmiðlar: www.dagur.net á völdum dagsetningum og www.sudurglugginn.is á völdum dagsetningum, fengið af vef með leyfi ritstjóra

Stúdentsprófsritgerðir í íslensku: frá Verslunarskólanum, ritgerðir nemenda sem gáfu leyfi, textar fengnir frá skólanum

Veftímarit: nokkur veftímarit, efni safnað af vefnum, leyfa aflað frá höfundum

Tölvupóstlistar: safnað af tveimur tölvupóstlistum með leyfi umsjónarmanna

Textavarp: af vefnum, valdar dagsetningar

Texti úr tónleikaskrám Sinfoníuhljómsveitar Íslands: með leyfi frá Árna Heimi Ingólfssyni

Kvikmyndadómar úr Morgunblaðinu: kvikmyndadómar Hjördísar Stefánsdóttur úr Morgunblaðinu

Safnaðarblöð: nokkur eintök af safnaðarblöðum Garðasóknar og Bessastaðasóknar, frá ritstjóra og með leyfi

MÍM – hreinsun texta

- Textarnir fengust í alls konar sniði
- Texti var dreginn úr pdf-skjölum (með sérstökum forritum eða skannaðir með ABBYY FineReader) og Word-skjölum, stundum fengust “hreinir” textar
- Textar voru hreinsaðir handvirkt og með forritum
- Mörg “erfið” tákn þurfti að laga, t.d. gæsalappir, úrfellingarmerki, bandstrik,…
- Notuð eru táknin « og » fyrir allar gæsir
- Tekin voru út efnisyfirlit, myndir, töflur o.s.frv.
- Í sumum textum þurfti að aðgreina fyrirsagnir
- Textum verður dreift í UNICODE (UTF8) stafatöflu…….

MÍM – Tilreiðsla, skipting í setningar, mörkun, lemmun, lýsigögn

- Notað var kerfið CorpusTagger sem var lýst fyrir MIM-GOLD fyrir tilreiðslu, mörkun og lemmun
- Nákvæmni mörkunar á textum MÍM hefur verið metin 88-95% eftir textum
- Lýsigögn (bókfræðilegar upplýsingar) voru skráð fyrir alla texta. Fyrir bækur var skráð:
  - Titill, höfundur(ar), f.ár höf., útgefandi, útgáfuár, ritstjóri (ef um það er að ræða)
  - Sambærilegt fyrir aðra textaflokka

MÍM – xml-snið til dreifingar texta

Textar verða aðgengilegir sem xml-skrár í sniði fyrir málheildir sem er skilgreint af Text Encoding Initiative (TEI). Bókfræðilegar upplýsingar, mörk og lemmur verða hluti af textunum. Ein setning úr bókinni *Með stein í skónum*:

```
<s n="1">
        <w lemma="hvernig" type="aa">Hvernig</w>
        <w lemma="varða" type="sfg2eþ">varðstu</w>
        <w lemma="ríkur" type="lkensf">ríkur</w>
        <c type="punctuation">?</c>
      </s>
```

MÍM: xml-snið – hluti af lýsigögnum (í “haus” skjalsins)

```
<sourceDesc>
      <biblStruct>
       <monogr>
        <author>Ari Kristján Sæmundssen</author>
        <title>
         <title type="main">Með stein í skónum  </title>
        </title>
        <imprint>
         <publisher>Salka</publisher>
         <date>2008</date>
        </imprint>
       </monogr>
      </biblStruct>
    </sourceDesc>
```

MÍM – aðgengi og notkun

Málheildin er aðgengileg á tvennan hátt:

1. Á vefsetri Stofnunar Árna Magnússonar í íslenskum fræðum er leitarbær útgáfa (http://mim.arnastofnun.is/)
2. Textar eru aðgengilegir sem xml-skrár af http://málföng.is/. Notendur sem vilja fá textana í tölvur sínar samþykkja notkunarleyfið
3. Á málfangasíðunni má einnig finna Excel-skjal með tíðnitölum nefnimynda

MÍM – Leitarkerfi

- Leitarkerfi MÍM er byggt á ***Glossa*** frá háskólanum í Osló (http://www.hf.uio.no/tekstlab/glossa.html)
- Í *Glossa* er notast við leitarvélina IMS Corpus Workbench (CWB) frá háskólanum í Stuttgart (http://cwb.sourceforge.net/)
  - Leitarmöguleikar hafa verið lagaðir að íslenskum mörkum
  - Vefviðmótið hefur verið þýtt
  - Leit gefur orðstöðulykil og hvaðan textinn er tekinn

Á http://mim.arnastofnun.is/ má einnig leita í málheild Orðtíðnibókarinnar og Fornritamálheildinni

MÍM – framtíðarmúsík

- Nýtt útlit á leitarsíðu hefur verið hannað en eftir er að setja það upp
- Bætt verður við möguleika til þess að velja texta til leitar
- Gera leitina notendavænni
- Notkunarleiðbeiningar (hjálp) verða endurgerðar
- N-stæður verða e.t.v. gerðar
- Talmál, Alþingisræður, texti og tal, verða gerðar aðgengilegar
- Merkingarfræðileg greining – tilraunaverkefni við að tengja gögn úr Orðanetinu við MÍM – hugsanlegt framhald
- Tengja MÍM við Islex og BÍN?

Textar frá öðrum málstigum

- Unnið er að gerð málheildar með textum sem eru fengnir úr Timarit.is (textar frá ca. 1870–1920)
- Skönnun er leiðrétt hálfhandvirkt með Skramba.
- Búið verður til annað lag af textunum með því að varpa textum yfir í nútímastafsetningu (með Skramba). Þá má nota tól fyrir mörkun og lemmun sem eru gerð fyrir nútímaíslensku.
- Gerð verður tveggja laga málheild sem verður aðgengileg fyrir leit.

Dæmi um notkun leitarkerfis MÍM

Orðtíðnibók

Orðmyndir af orðinu „hestur“ til þess að skoða útlitið
Finna tvö fornöfn í röð, eða fn fn fn ao……

.

MÍM

það-leppur (« Það er ekkert víst að hann sé hérna , »)
lýsingarorð með „kjóll“
setja/koma á fót (setja á fót/stofn)
klæðast flík (klæðast +(lo no))
e-ð felst í e-u
horfa í gaupnir sér (líkamshluti og afturbeygt fornafn (persónufornafn)
(Reynið að leysa þessi verkefni/SH)

Dæmi um notkun leitarkerfis MÍM

MÍM (frh.)
so. þgf. þf. (ryðja/brjóta sér leið)

MÍM (blogg)
mæ
nýja þolmyndin (það var barið mig)

Dæmi um notkun leitarkerfis MÍM

MÍM (frh.)

Bjarki Karlsson (27.9.2014) af Facebook

Þó öfgasveitir ekki beint ég styðji
og aldrei le***pji*** te úr þeirra bolla
þá finnst mér sætt að farísear biðji
fyrir þeim sem heimta af okkur tolla.

Leita að „pji“ inni í orði

(Eitt dæmi:
Þú veist maður skilur að krakkarnir glepjast **glepjist** þegar maður gerir það sjálfur ..
Úr talmáli (Ístal))

(hugmynd ER)

Dæmi um notkun leitarkerfis MÍM

MÍM (frh.)

Káinn (úr Vesturfaraþætti Egils Helgasonar)

Þetta er ekki þjóðrækni
og þaðan af síður guðrækni
heldur íslensk heiftrækni
og helvítis bölvuð langrækni.
Finna orð sem enda á -rækni

Dæmi um notkun leitarkerfis MÍM

Fornrit

sögn með sverð

lýsingarorð með skjöldur

brjóta sér leið, ryðja sér braut

(…hún horfin úr þeim farveg og hafði **brotið sér nýjan** farveg austur um sandana . Kerling ein..)

mér líkur/líkur mér

Notkun MÍM í máltækniverkefnum

BÍN – Notuð til þess að auka orðaforða
BÍN (Kristín Bjarnadóttir)
Skrambi – Jón Friðrik Daðason hefur notað *MÍM* og Íslenskan orðasjóð við gerð Skramba
Valdir textar notaðir við gerð „word prediction“ fyrir lesblinda

……

Íslenskur orðasjóður og Skrambi

Leiðréttingarforritið Skrambi skilar uppástungum í líkindaröð

Líkindi hverrar uppástungu er margfeldi tveggja þátta:

- Líkindin á orðinu sem stungið er upp á
- Líkindin á villunni (t.d. y→i)

Líkindi orða eru metin út frá tíðni þeirra í Íslenskum orðasjóði

Listi yfir 5.000 algengustu stafsetningarvillurnar í Íslenskum orðasjóði (ásamt tíðni þeirra og leiðréttingum) er notaður til að meta líkindin á mismunandi villum (Jón Friðrik Daðason)

MÍM og Skrambi

Villur eru sagðar vera samhengisháðar ef þær eru til í málinu

- Vorið er á næsta **leyti**
- **Mér** langar heim

Skrambi notar hluta af MÍM til þess að búa til reglur til að leiðrétta samhengisháðar villur

- Textar með tiltölulega mikið af villum (t.d. blogg) eru ekki notaðir

Dæmi um reglur:

- Breytum **leyti** í **leiti** ef *persónufornafn* kemur fyrir beint á undan
- Breytum **sína** í **sýna** ef beint á eftir koma orðmyndirnar **fram** og því næst **á** (Jón Friðrik Daðason)

Gestir

Kristín Bjarnadóttir: MÍM í BÍN

Kristján Rúnarsson: N-stæðu skoðari