

Mörkuð íslensk málheild

Sigrún Helgadóttir

X loka
Fjöldi niðurstaðna: 20
Síða: 1
Tilni | Raða hjálp

TIMARIT-TO8	að . Annað þeirra er Mörkuð íslensk	málheild	(MÍM) (Verkefnið er unnið
TIMARIT-TO8	hvers kyns sófn rafrænna texta en hugtakið	málheild	(d. korpus , e. corpus)
TIMARIT-TO8	Conrad Rippen lýsa muninum á textasafni og	málheild	bannig : « A corpus is not
TIMARIT-TO8	að tiltekið safn rafrænna texta geti kallast	málheild	þarf það m.ö.o. að uppfylla ákveðin skilyrði
TIMARIT-TO8	Ekkert ofangreindra safna getur þó talist vera	málheild	í þeim skilningi sem hér er lagður
TIMARIT-TO8	hefur verið nefnt , getur líka kallast	málheild	í þeim skilningi að það nær til
TIMARIT-TO8	(þótt hvorugt þeirra geti talist fullgild	málheild	m.t.t. áðurgreindra viðmiða) . Málið vandast
TIMARIT-TO8	utan um í heild sinni , t.d.	málheild	með íslensku ritmáli á 20. öld eða
TIMARIT-TO8	niðurstöðum rannsókna sem byggðar eru á tiltekinni	málheild	er að samsetning hennar endurspeglir raunverulega málnotkun
TIMARIT-TO8	hægt er að fella efniviðinn inn í	málheild	eða rannsaka tiltekin einkenni sem þar birtast
TIMARIT-TO8	talmálsefni , t.d. sem hluta af almennri	málheild	, kæmi að góðu gagni við orðabókagerð
TIMARIT-TO8	Sé stuðst við nægilega stóra og fjölbreytilega	málheild	gefur hún líka mikilvæga vitneskju um tíðni
TIMARIT-TO8	að hafa beina tengingu úr orðabókargreinunum í	málheild	eða textasafn bannig að notendur geti sjálfir
TIMARIT-TO8	er verið að gera með Markaðri íslenski	málheild	(MÍM) . Þá er hægt
TIMARIT-TO8	varðar viðhald og eflingu slíkra safna .	Málheild	sem er ætlað að endurspegla samtímamálið úrelidist
TIMARIT-TO8	inn og taka út . Þótt miðlæg	málheild	sé til og öllum opin getur hún
TIMARIT-TO8		lid	. Slíkar málheildir eru sums staðar til
TIMARIT-TO8		lid	sé ekki til hefur mikil árangur náðst
TIMARIT-TO8		lid	þýsks ritmáls með leitarmöguleikum) á Veraldarvefnum
TIMARIT-TO8		lid	(sbr. Sigrún Helgadóttir 2004) .

Textasafn: Tímarit
Titill: Talmál og málheildir – talmál og orðabækur
Höfundur: Ásta Svavarsdóttir
Ritstjóri: Guðrún Kvaran
Útgefandi: Stofnun Árna Magnússonar í íslenskum fræðum
Í: Orð og tunga
Ár: 2007

MÍM

- Uppruni verkefnisins
- Hvað er mörkuð málheild?
- Hvernig nýtist málheildin?
- Efnisöflun fyrir MÍM
- Textaflokkar
- Birting MÍM
- Opnun

Uppruni verkefnisins

Mörkuð íslensk málheild (MÍM) er eitt af verkefnum sem var styrkt af tungutækniverkefni menntamálaráðuneytisins (2000-2004)

Verkið hófst árið 2004 og hefur verið fóstarað af Orðabók Háskólans og síðar Stofnun Árna Magnússonar í íslenskum fræðum

Önnur fjármögnun:

- Norræna ráðherranefndin (Nordisk Netordbog)
- Rannís (Hagkvæm máltækni utan ensku – íslenska tilraunin)
- Nýsköpunarsjóður námsmanna (nokkrir styrkir)
- Rannsóknarsjóður Háskólans (nokkrir styrkir)
- META-NORD

Hvað er mörkuð málheild?

Mörkuð málheild (e. *tagged corpus*)

- Safn fjölbreyttra tölvutækra texta sem hafa verið greindir á málfræðilegan hátt
- Hverjum texta fylgja upplýsingar um textann sem búturinn er úr
- Hverri orðmynd fylgja þær málfræðilegu upplýsingar sem málheildin geymir
- Málheildin er skráð í stöðluðu sniði

Hvernig nýtist málheildin?

- Úr málheildinni má lesa gagnlegan fróðleik um:
 - tíðni orðflokka, orða og beygingarmynda, orðasambönd, setningargerð, merkingu o.fl.
- Nýtist t.d. við gerð:
 - orðabóka, leiðréttingarforrita, þýðingarforrita, búnaðar fyrir talgreiningu og talgervingu og gerð hjálparforrita fyrir blinda, heyrnarskerta, hreyfihamlaða og þá sem glíma við skriftar- og lestarörðugleika svo og við kennslu

Efnisöflun fyrir MÍM

- Safna skyldi fjölbreyttum textum með um 25 milljónum lesmálsaða af frumsömdum textum, rituðum á árunum 2000–2009 eftir höfunda sem hafa íslensku að móðurmáli
- Safnað var textum sem voru aðgengilegir í rafrænu formi
- Leyfis var aflað hjá rétthöfum til þess að fá að nota texta varða af höfundarrétti
- Textar úr útgefnum bókum voru styttnir um 20%

Textaflokkar í Markaðri íslenskri málheild (MÍM)	Fj. skráa	Fjöldi orða	%
Textar úr prentuðum bókum	168	5.972.893	23,89
Dagblöð, prentuð og af vef	12.725	5.779.509	23,12
Opinberir textar (skýrslur, dómar, frumvörp, lög, ræður þingmanna)	1.246	3.513.990	14,06
Tímarit (prentuð og rafræn)	311	2.501.222	10,00
Blogg	8.998	1.976.706	7,91
Pistlar af Vísindavef	4.949	1.838.909	7,36
Texti af vefsetrum fyrirtækja, samtaka og stofnana	106	1.337.764	5,35
Upplesið efni	1.196	694.506	2,78
Nemendaritgerðir	51	666.042	2,66
Talmál	4	504.318	2,02
Óflokkað	46	214.663	0,86
Samtals	29.800	25.000.522	100,00

Birting MÍM

Mörkuð íslensk málheild verður aðgengileg á tvennan hátt:

1. Á vefsetri Stofnunar Árna Magnússonar í íslenskum fræðum er leitarsíða (<http://mim.hi.is/>)
2. Sækja má textana í xml-sniði í fyrir málheildir í gegnum síðuna **<http://www.málföng.is/>** gegn því að skrifa undir notkunarleyfi

<http://www.málföng.is/> er vefsvæði þar sem má finna margvísleg íslensk málföng (Language Resources)

Mörkuð íslensk málheild

Mennta- og menningarmálaráðherra

Katrín Jakobsdóttir

opnar nú málheildina

Mörkuð íslensk málheild