Stofnun Árna Magnússonar í íslenskum fræðum

Orðfræðisvið:
Kynning á gögnum

Heimsókn íslenskunema
25.9.2012

Kristín Bjarnadóttir

STOFNUN ÁRNA MAGNÚSSONAR
Í ÍSLENSKUM FRÆÐUM

Viðfangsefni á orðfræðisviði

- Mál og málnotkun - einkum orðfræðileg viðfangsefni (orð, orðasambönd, orðaforði)
- Orðabækur og orðabókagerð - fræðileg og hagnýt orðabókafræði
- Máltækni – uppbygging málsafna og þróun máltæknitóla

Starfsemi á orðfræðisviði

- Rannsóknir og þróunarstarf
- Hagnýt verkefni
  - Orðabókagerð
  - Uppbygging málgagna
  - Þróun máltæknitóla

Starfsemi á orðfræðisviði

- Varðveisla, efling og úrvinnsla málsafna
  - Seðlasöfn Orðabókar Háskólans
  - Rafræn textasöfn og málheildir
  - Ýmis gagnasöfn
- Miðlun þekkingar
  - Kennsla og leiðsögn
  - Útgáfustarf og upplýsingagjöf
  - Aðgangur að gagnasöfnum

Gagnasöfn: Hvað finnst hvar?

- ÍSLEX: Íslensk-skandínavísk orðabók → Þórdís Úlfarsdóttir
- Mörkið íslensk málheild: → Sigrún Helgadóttir
- Ritmálssafn Orðabókar Háskólans
- Talmálssafn Orðabókar Háskólans
- Íslenskt textasafn
- Beygingarlýsing íslensks nútímamáls
- Íslenskt orðanet

Ritmálssafn Orðabókar Háskólans

http://www.arnastofnun.is/page/arnastofnun_gagnasafn_ritmal

- Tímabilið: 1540 til u.þ.b. 1980/1985
- U.þ.b. 700 þúsund uppflettiorð
- 2,5 milljónir dæma úr textum, mestmegnis af prenti
- Orðtekið með hefðbundnum hætti, með lestri

Í Ritmálssafninu (Rms.) er gríðarlegur orðaforði en eini aðgangurinn er í gegnum uppflettiorðið.

Í Rms. er greiður aðgangur að aldursmerkingu orða en safnið er ekki heppilegt til neins konar tíðnirannsókna.

Talmálssafn Orðabókar Háskólans

- Seðlasafn með umsögn heimildarmanna um talmál
- Efninu var safnað á árunum 1956-2005, í tengslum við vikulega útvarpsþætti yfir veturinn
- Svör bárust í bréfum og símtölum
- Seðlarnir eru merktir heimildamönnum og þar kemur fram hvaðan af landinu þeir eru
- Þeir eru því góð heimild um staðbundið málfar

Safnið er enn sem komið er
aðeins aðgengilegt á seðlum.

Íslenskt textasafn

http://www.lexis.hi.is/corpus/leit.pl

Orðstöðulykill, u.þ.b. 65 milljón lesmálsorð, úr samfelldum textum, frá fornu máli til nútímans.

Safninu fylgir ekki orðalisti og engin leið er að komast að því hver orðaforðinn er.

Textasafnið er frábær uppspretta dæma.

Flokkun dæma er mjög seinleg, m.a. vegna **tvíræðni** orðmynda, og vegna þess hve dæmin geta orðið mörg.

Beygingarlýsing íslensks nútímamáls (BÍN): bin.arnastofnun.is

U.þ.b. 270 þúsund beygingardæmi úr nútímamáli.

Reynt er að birta sem afbrigði eftir því sem nokkur kostur er.

Ofan við beygingardæmin eru athugasemdir um notkun, t.d. þegar afbrigði eru bundin við tiltekna merkingu eða notkunarsvið.

BÍN er notuð í máltækni og verkið er ekki stafsetningarorðabók.

BÍN er í vinnslu og athugasemdir, leiðréttingar og ábendingar um viðbætur eru þegnar með þökkum!

Gögnin er hægt að sækja og nota skv. skilmálum á vefsíðunni.

Tölur um tvíræðni orðmynda í BÍN

Orðmyndir (= strengur, án greiningar (marks))
2,8 milljónir orðmynda
1,8 milljón orðmynda með aðeins einu marki
1 milljón tví- eða margræðar orðmyndir

Tvíræðni beygingarmynda í BÍN
Beygingarmyndir (= orðmynd með marki)

Beygingarmyndir alls	5.881.374	
Einræðar beygingir	1.850.090	31,5\%
Tvíræðar* í einni flettu	3.619.482	61,5\%
Tvíræðar milli flettna	63.641	1,1\%
Tvíræðar innan og milli flettna	348.161	5,9\%

* tvíræður = tví- eða margræður

Í ómörkuðum texta er tvíræðnin verulegt vandamál.

Íslenskt orðanet

http://www.ordanet.is

Jón Hilmar Jónsson og Þórdís Úlfarsdóttir

Sundurgreining og flokkun orðaforðans eftir merkingareinkennum.

,,Grunnhugmyndin að baki verkefninu er sú að rekja megi merkingarvensl innan orðaforðans út frá setningarlegum og orðmyndunarlegum venslum á milli orða.”

STOFNUN ÁRNA MAGNÚSSONAR
Í ÍSLENSKUM FRÆÐUM

Önnur gagnleg gögn:

Ritmálssafnið nær ekki aftur fyrir 1540. Til samanburðar:

Fornmálsorðabók: ONP (Ordbog over de **norrøne prosasprog)**
http://www.onp.hum.ku.dk/

Stór textasöfn, ómörkuð:

Tímaritavefur Landsbókasafns-Háskólabókasafns: **timarit.is**
Aldursmerkingar eru mjög þægilegar en leitin er seinleg**.**

Íslenskur orðasjóður, hjá Háskólanum í Leipzig.
http://wortschatz.uni-leipzig.de/ws_ice/
250 milljón lesmálsorð af íslenskum vefsíðum.

ÍSLEX

islex.hi.is

Mörkuð íslensk málheild

mim.hi.is

STOFNUN ÁRNA MAGNÚSSONAR
Í ÍSLENSKUM FRÆÐUM

Takk fyrir áheyrnina.

Kristín B: kristinb [hjá] hi.is

STOFNUN ÁRNA MAGNÚSSONAR
Í ÍSLENSKUM FRÆÐUM