**Kristín Bjarnadóttir**

# Hvert á að sækja orðaforðann í orðabók?

**Málþing Orðs og tungu 2012.**
Íslenska sem viðfangsmál í íslensk-erlendum orðabókum.
Sjónarmið við aðferðir og öflun, val og framsetningu efnisins.

4. maí 2012

STOFNUN ÁRNA MAGNÚSSONAR
Í ÍSLENSKUM FRÆÐUM

# Efnisskipan

Hér er sjónum eingöngu beint að **flettiorðalistanum** sjálfum.
Í erindi Jóns Hilmars verður fjallað um aðföngin í víðara samhengi og áhrif á efnistök í orðabókartextanum.

- Kostir og gallar **hefðbundinna aðferða** við að búa til flettiorðalista í íslenskar orðabækur.

- Orðtaka úr **rafrænum textum**, með hefðbundnu sniði (dæmaleit).

- **Vélræn orðtaka**, úr völdu efni.

- Orðaforðinn í **Markaðri íslenskri málheild** (MÍM), m.t.t. orðtöku til nota í íslenskum orðabókarstofni.

# Hvaðan er orðaforðinn í íslenskum orðabókum?

**Erfðaefnið**: Útgefnar orðabækur
Orðalistar
Seðlasöfn og önnur orðfræðileg gagnasöfn

**Hefðbundin orðtaka:** Fyllt upp í augljós göt, eftir efnisflokkum, úr tiltækum prentuðum ritum.

**Orðtaka úr rafrænum textum:** Íslenskt textasafn, tímarit.is o.fl.
Nýtist einkum til dæmaleitar.

**Vélræn orðtaka:** Úr tilfallandi efni, t.d. daglega úr dagblöðum.
Reynt að fanga allan orðaforðann í viðkomandi texta.

**Málheildir**: Fullgreint skipulegt safn af textum: **MÍM**
Allur orðaforðinn í heimildunum fangaður og flokkaður eftir tíðni.

# Erfðaefnið

**Meginrit**: *Íslensk-dönsk orðabók* Sigfúsar Blöndals (1920-1924)
*Íslensk orðabók* (1963, 1983 o.áfr.)

**Gallarnir** við að nota orðalista úr eldri orðabókum og söfnum:

- Skilgreining á markhópi og yfirgripi á ekki við í nýju verki.
- Hluta orðaforðans vantaði til skamms tíma í orðabækur.
- Tilhneiging er til að safna sjaldgæfum orðum, á kostnað þess sem algengt er (sbr. Ritmálssafn OH).
- Orðaforðinn úreldist og því þarf bæði að bæta við og grisja.
- Villur og draugorð geta átt sér langa ævi (sjá dæmi →).

**Kostir** við að nota orðalista úr eldri orðabókum og söfnum:

- Það sparar tíma.
- Orðaforðinn er gríðarlegur (t.d. í samanburði við MÍM).

# Dæmi um draugorð í orðabókum

**nösgæs** KV *e.k. gæs eða ef til vill önd. (*Ío* 1963, 1983)

Orðið er í gátu í Heiðreks sögu og merkir ‘andardráttur’. Lausnin er í sögunni sjálfri en samt sem áður hefur þetta orð komist á flug í orðabókum og úr því orðið raunverulegur fugl:

> NÖSGÁS, f., **anser** nasutus (nös, gás), **anas** ...
> forte = toppönd mergus fuscus cirratus. (*Lexicon Poeticum antiquæ linguæ septemtrionalis.* Hafniæ 1860)
> [Leturbreyting KB; *anser* ‘*gæs*’, *anas* ‘*önd*’]

Sjá grein Aðalsteins Eyþórssonar, 1998: Síðasti furðufuglinn – eða gæsin sem stóð á öndinni. *Frejas Psalter til brug for Jonna Louis-Jensen.* København. Bls. 5-7. Birt í Kistunni.

*Fæst dæmi um óþarft efni í orðabók eru svona dramatísk!*

# Dæmi um orðaforða fyrir tvímálaorðabók

**Norræna verkefnið** (1994-1997): Íslenskur orðabókarstofn

Upprunalegur orðalisti var unninn upp úr gögnum OH, seðlasöfnum, orðalista úr *Íslenskri orðtíðnibók* og efni frá Jóni Hilmari Jónssyni. Textasafn OH var einnig notað. Listarnir voru bornir saman við prentaðar orðabækur, t.d. *Íslenska orðabók* (1983).

Allt þetta efni er hefðbundið orðabókarefni, **erfðaefnið**, nema orðaforðinn úr *Íslenskri orðtíðnibók* (30 þúsund flettur). Þar er **allur orðaforðinn** úr 500 þúsund lesmálsorðum.

Orðaforðinn í Norræna verkefninu:  241.784 flettur.

Allur orðaforðinn var flokkaður eftir orðgerð og flettugildi, t.d. í lexíkalíseraðar og virkar samsetningar.

# Hefðbundin orðtaka í Norræna verkefninu

Efni um ,,ýmsa þætti í nútímaþjóðfélagi“ var orðtekið:

- Bæklingar um heilbrigðismál, félagsmál, tryggingar, bankakerfið, sálfræði, lögfræði fyrir almenning, sjávarútveg og tölvur.

- Námsbækur í örverufræði, matvælafræði, vélsmíði, húsgagnasmíði, íþróttum o.fl.

- Tímarit og Morgunblaðið (sem reyndist einna drýgst).

Umfangið: Nokkur þúsund orð, 1994 u.þ.b. 2.500 orð, ýmist af prenti eða úr rafrænu efni.

Sigurborg Hilmarsdóttir stjórnaði þessum orðtökunni og við hana unnu t.d. 4 starfsmenn sumarið 1995.

# Dæmi um orðaforða úr Morgunblaðinu úr orðtökunni fyrir Norræna verkefnið

áherslupenni
barnaís
blekksprautuprentari
buxnadragt
eldislax
fjárlagahalli
flísefni
grasrótarhreyfing
heimasíða
húsbréfalán
kapalkerfi
krókaleyfi
leikskólakennari
líkamsklukka
ljósleiðari
margmiðlun
myndlykill
raðgreiðsla
ríkisvíxill
talhólf

Orðtakan var mjög tímafrek og var unnin með hefðbundnum aðferðum, þ.e. leitað var að orðum sem þóttu hafa gildi í orðabók. Virkar samsetningar voru ekki teknar með.

# Dæmasöfnun úr rafrænum textum

**Textasafn** OH: Orðstöðulykill, u.þ.b. 60 milljón lesmálsorð.

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

Orðabókarmaðurinn þarf því að vita fyrir fram
að hverju hann er að leita!

Textasafnið er frábær uppspretta dæma og sama á við um timarit.is.

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

## Tölur um tvíræðni orðmynda í Beygingarlýsingunni (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.

## Vélræn orðtaka: Orðtökutól í BÍN

Hluti af BÍN er orðtökutöl.
Það á að finna **orðmyndir** í texta sem ekki eru í BÍN.
Tólið er á enn á tilraunastigi og er frumstætt.

Til þess að orðtökutól komi að gagni þarf
- samanburðarlista (t.d. úr BÍN eða því verki sem unnið er að)
- máltæknibúnað:
    - lemmun
    - orðgreiningu (samsetningu o.þ.h.)
    - síu fyrir rusl

- sjálfvirka mötun úr tilteknum textum, t.d.
    - af vefnum
    - úr dagblöðum, o.s.frv.

... t.d. til þess að finna **öll ný orð** í dagblaði þann daginn ...

# Dæmi um vélræna orðtöku: Kynning á erindi JHJ

Gerð íslensk-erlendrar orðabókar á nútímavísu krefst mikils og margbreytilegs gagnaefnis sem myndar undirstöðu fullbúins orðabókartexta. Orðabókartextinn mótast við mat á þeim gögnum sem fyrir liggja og eftir atvikum nánari greiningu á þeim. Meðferð og nýting gagnanna helst svo í hendur við það hlutverk sem orðabókinni er ætlað að gegna.
Í íslensk-erlendri orðabók er íslenska í forgrunni sem viðfangsmálið og í stórum dráttum má velja og skipa efni hennar óháð einingum markmálsins sem viðfangsmálið kallar á. Aðföng slíkrar orðabókar eru því að miklu leyti hin sömu og þörf er á við gerð einmála íslenskrar orðabókar. Efnisval og afmörkun markast hins vegar af tilliti til þarfa og forsendna erlendra notenda jafnframt því sem greiða þarf íslenskum notendum leið að orðum og orðanotkun í markmálinu.
Flettiorðin mynda hinn ytri ramma orðabókartextans og val þeirra er í brennidepli. Önnur efnisatriði hverfast um flettiorðin og eru ekki í sama mæli áþreifanleg sem sjálfstæðar einingar. Það á m.a. við um orðasambönd og merkingarlegt samhengi jafnt orða sem orðasambanda. Aðgangur notenda að þessum innri þáttum er oft harla ógreiður og hlutur þeirra rýrari en efni standa til.
Þessar aðstæður hafa löngum einkennt orðabækur í prentuðum búningi þar sem textinn kemur fram í föstum skorðum. Í rafrænni orðabók má losa um þær skorður og draga athyglina í ríkari mæli að hinum innri efnisþáttum og innbyrðis venslum orðabókareininganna, jafnt formlegum sem merkingarlegum. Með því gefst einnig kostur á fjölþættari samanburði málanna tveggja óháð stöðu þeirra sem ...
(*Af vefsíðu um málþing Orðs og tungu)*

## Útkoman úr orðtökutólinu

*og gagnaefnis fyrir eftir nýting svo í að Í einmála af til jafnframt um m.a rýrari en þar fram orðabókareininganna Með einnig yfir hvernig nýta orðabókargögn nýjar nýsköpun orðabókargagna*

• Orðtökutólið á að finna allar orðmyndir (strengi) sem ekki koma fyrir í BÍN: gagnaefnis, einmála, orðabókareininganna, orðabókargögn, orðabókargagna (4 uppflettiorð).

• Orðtökutólið á ekki að finna orð sem eru í BÍN: nýting, rýr, nýta, nýr, nýsköpun

• Óbeygjanleg orð eru ekki enn í samanburðarlistanum: og, fyrir, eftir, svo, í, að ...

• Orðtökutólið lemmar ekki og skilar býsna óaðgengilegum listum af orðmyndum (og öðrum strengjum). Úrvinnslan er seinleg.

## Hefðbundin orðtaka, vélræn orðtaka, málheild

**Hefðbundin orðtaka** byggist á lestri. Efnið er **flokkað** jafnóðum.

**Vélræn orðtaka** byggist á því að finna orðmyndir sem ekki eru í samanburðarefni. Orðtakan skilar öllu efni **óflokkuðu**.

Með orðtöku úr markaðri og lemmaðri málheild gefst í fyrsta sinn kostur á skilvirkri vinnu við **orðalista með tíðnitölum**. Tölurnar er hægt að nota til **flokkunar**.

Fyrsta orðabókin sem byggist alfarið á orðtíðni og málheild er

*Collins Cobuild English Language Dictionary.* 1987.
*“over 70,000 references”*

Hún er byggð á málheild með u.þ.b. 26 milljón lesmálsorðum.
Í MÍM verða u.þ.b. 25 milljón lesmálsorð.

## Mörkuð íslensk málheild, MÍM

Í MÍM verða 25 milljón lesmálsorð. Textarnir eru allir frá 21. öld og þeir eru úr eins fjölbreyttu efni og nokkur kostur er.

Hverju orði fylgir mark, þ.e. orðflokkur og beygingarlegar upplýsingar.

Markamengið er stórt, yfir 600 mörk.

Lemma fylgir hverju orði og þar er lykill að orðaforðanum fyrir orðabókarmenn.

Mörkun og lemmun á MÍM er ekki lokið og tölur um lemmufjölda og tíðni liggja ekki fyrir enn.
Tíðnin mun gagnast orðabókarmönnum vel!

Ritstjóri MÍM er Sigrún Helgadóttir.

## Orðamyndafjöldinn í MÍM, bráðabirgðatölur

Athugun á orðmyndum í 17,5 milljónum lesmálsorðum úr MÍM, í samanburði við beygingarmyndir í BÍN:

| | |
|---|---|
| Tókar: | 16.245.429 |
| Orðmyndir (með marki): | 737.856 |
| Þar af í BÍN: | 425.238 |
| Þar af ekki í BÍN: | 312.618 |

Orðmyndir úr BÍN eru taldar vera tækar íslenskar orðmyndir. Orðmyndir sem ekki voru í BÍN voru skoðaðar hver fyrir sig.

Hlutfall “aðskotaefnis” í MÍM, þ.e. efnis sem ekki telst vera íslenskur orðaforði er u.þ.b. 13% (af 738 þúsund. Gróf ágiskun.)

## Orðmyndir (strengir) í MÍM sem ekki eru í BÍN: Hvað leynist í textunum?

60% íslenskar orðmyndir (af 312 þúsund)
40% annað:
- 25% útlenska
- 6% villur (ritvillur)
- 1,7% skammstafanir og styttingar
- 0,7% tölvumálsstrengir (vefföng, slóðir ...)

Allar orðmyndir sem hafa verið aðlagaðar að íslensku málkerfi flokkast sem ,,íslenskar orðmyndir“, þ.m.t. allar beygðar slettur (*pródúsjónin*) og beygð erlend nöfn, t.d. grísku nöfnin *Parmenídes (Parmenídesar) og Erastoþenes (Erastoþenesar)*.

Í þessum hluta MÍM eru sennilega 125 þúsund orð (lemmur) sem ekki eru í BÍN. Samanlagður lemmufjöldi í BÍN og MÍM gæti verið yfir 400 þúsund.

*Athugið! Orðaforðinn í BÍN er úr orðalistum.*
*Hann er ekki flokkaður eftir flettugildi í orðabók.*

# Orðmyndir í MÍM sem ekki eru í BÍN

| | *Orðmynd* | *Mark* | *Lemma* | *Ofl.* | |
|---|---|---|---|---|---|
| 1262 | vísindavefnum | nkeþgs | vísindavefur | kk | |
| 282 | sonu | nkfo | sonur | kk | |
| 258 | grunnskólabekk | nkeþ | grunnskólabekkur | kk | |
| 231 | r-listans | nkeegs | r-listi | kk | ← |
| 177 | knattspyrnustjóri | nken | knattspyrnustjóri | kk | |
| 158 | skjöldu | nveo | skjölda>skjöldur | kvk>kk | |
| 151 | vísindavefsins | nheegs | vísindavefur* | hk>kk | |
| 141 | ameríkufari | nken-s | ameríkufari | kk | |
| 126 | íbúðalánasjóðs | nkee-s | íbúðalánasjóður | kk | |
| 122 | r-listinn | nken-s | r-listinn | kk.mgr. | ← |
| 111 | vopnfirðinga | nkfe-s | vopnfirðingur | kk | |
| 101 | ríkislögreglustjóra | nkee | ríkislögreglustjóri | kk | |
| 99 | óskarinn | nkeng | óskar | kk | |
| 96 | madur | nken | **madur** | kk | |
| 95 | mogganum | nkeþgs | moggi | kk | |
| 93 | danakonungur | nken-s | danakonungur | kk | |
| 93 | danakonungs | nkee-s | danakonungur | kk | |
| 84 | aðalleikendur | nken-s | aðalleikandi* | kk | |

Lemmunin í MÍM er ekki leiðrétt. Í henni eru villur, bæði í uppflettimyndum og orðflokkum. Hér eru hástafir teknir út til einföldunar í keyrslum.

## MÍM og Cobuild: Hvernig yrði orðabók úr MÍM?

Birmingham-málheildin sem Cobuild-orðabókin er byggð á er á stærð við MÍM.

Tilgáta mín er að orðaforðinn úr MÍM myndi skila gloppóttri orðabók.

Ég tel erfðaefnið í orðabókum líka vera nauðsynlegt!

Tíðnitölurnar skila e.t.v. ekki eins afgerandi niðurstöðum fyrir íslensku og fyrir ensku og ein af ástæðunum gæti t.d. legið í mismunandi reglum um samsett orð:

kransæðarhjáveitugræðlingur
coronary artery bypass graft

Þarna er eitt orð í íslensku fyrir fjögur í ensku.
E.t.v. þurfum við líka að telja orðhluta!

## Niðurstaða

Eftir þennan stutta samanburð á mismunandi aðferðum við að velja orðaforða í íslenska orðabók er niðurstaðan þessi:

- Hefðbundnir orðalistar eru enn sem komið er ómissandi. Þá þarf að grisja og uppfæra með nýjum aðferðum.

- Vélræn orðtaka er ekki enn komin í gagnið en mun nýtast til að hafa fingurinn á púlsinum og finna ný orð.

- MÍM mun verða gríðarlega mikilvæg heimild um orðaforðann. Til þess að hún komi að fullum notum þarf að greina samsett orð í sundur.

- Engar vélrænar aðferðir eru enn til til þess að meta flettugildi orða. Þar þarf orðabókarmaðurinn að meta efnið á hefðbundinn seinlegan hátt, þótt hægt sé að styðjast við tíðnitölur.

- Loks má nefna að í tvímálaorðabók setur markmálið svip sinn á orðaforðann.

**Takk fyrir áheyrnina.**

**Kristín B: kristinb [hjá] hi.is**

STOFNUN ÁRNA MAGNÚSSONAR
Í ÍSLENSKUM FRÆÐUM

já - er svarið

# Beygingarlýsing íslensks nútímamáls

Stofnun Árna Magnússonar í íslenskum fræðum



## köttur

Karlkynsnafnorð

| | Eintala | | Fleirtala | |
|---|---|---|---|---|
| | án greinis | með greini | án greinis | með greini |
| Nf. | köttur | kötturinn | kettir | kettirnir |
| Þf. | kött | köttinn | ketti | kettina |
| Þgf. | ketti | kettinum | köttum | köttunum |
| Ef. | kattar | kattarins | katta | kattanna |

vera Leita

Leit að beygingarmynd