Íslenskar málheildir
einmála og margmála

Málstofa ELRC
Gagnagrunnar og vélrænar þýðingar
28. september 2018

Steinþór Steingrímsson

STOFNUN ÁRNA MAGNÚSSONAR
Í ÍSLENSKUM FRÆÐUM

Einmála málheildir

- Orðtíðnibókin (1991) 500 þúsund orð
- MÍM (2012) 25 milljón orð
- Risamálheildin (2018) 1250 milljón orð

Samhliða málheildir

- Engar íslenskar til
- Íslenska í stöku margmála málheildum
- Of lítið efni til að þjálfa góðar þýðingarvélar

Samhliða málheild fyrir vélþýðingar

- **Hófst í fyrra**
  - Styrkt af Menntamálaráðuneytinu og Máltæknisjóði

- **Markmið**
  - Kortleggja hvað er til af efni fyrir íslensku
  - Safna því saman sem er gefið út með opnu leyfi
  - Bæta við ef mögulegt er
  - Para saman önnur tvímála gögn sem til eru og eru aðgengileg

Hvað var til?

- **OPUS**
  - Skjátextar (1400 þúsund setningar)
  - Stýrikerfisþýðingar (KDE, Gnome, Ubuntu)
- **Tilde MODEL – EMA** (420 þús.)

- **EES þýðingar**
  - En reyndar bara í skjölum sem ekki hafa verið pöruð saman
- **Fréttir frá ESO, bækur úr þýðingarrétti o.fl.**

Stækkuðum og hreinsuðum

- Bjuggum til aðferð til að meta gæði samhliðunarinnar.
- Notuðum hana til að hreinsa gögnin

	Fjöldi	Rangt	Gallað
OpenSubtitles	1.260.000	3,13%	0,75%
KDE/Ubuntu	60.000	4,50%	4,50%
EMA	400.000	2,00%	0,50%

EES-þýðingar

- Þýðingar á lögum, reglugerðum og öðrum opinberum skjölum sem tengjast EES

- Textar frá því um aldamót til 2017

	Fjöldi	Rangt	Gallað
EES	1.900.000	2,75%	6,25%

Tilraunir með annað

- **Bækur**
  - Biblían
  - Fornrit
  - Skáldsögur

- Skrapa samhliða texta af netinu

Bráðabirgðaniðurstaða

	Ísl. orð	Línur	Rangt	Gallað
Biblían	670.000	65.000	0,00%	0,00%
Skáldsögur	197.000	12.000	0,50%	3,00%
Fornsögur	270.000	17.000	3,50%	7,50%
EES	23.400.000	1.900.000	2,75%	6,25%
ELRC	29.000	2.200	4,50%	0,50%
EMA	3.800.000	400.000	2,00%	0,50%
ESO	250.000	12.000	0,00%	0,50%
OpenSubtitles	6.800.000	1.260.000	3,13%	1,75%
Tatoeba	55.000	8.200	0,00%	0,00%
Ubuntu	40.000	10.000	2,00%	0,00%
KDE4	270.000	50.000	4,50%	4,50%
	35.511.000	**3.686.400**		

Hvar annars staðar eru gögn?

- Meira af því sama (takmarkað magn)
- Vefsíður á mörgum tungumálum
- Þýddar bækur (leyfismál geta verið flókin)
- Þýðingarminni (þýðingar fyrir hið opinbera)
- Fleira?

Aðgengi að málheild

- **Málheildarkerfi**
  - sambærilegt við einmála málheildirnar
- **Málföng.is**
  - Hægt að hlaða gögnunum niður á stöðluðu sniði