

Det islandske ordklasseopmærkede korpus

MÍM

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 söfn rafræna 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 rafræna 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 þannig 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 úreldest |
| TIMARIT-TO8 | inn og taka út . Þótt miðlæg | málheild | sé til og öllum opin getur hún |
| TIMARI | Textasafn: Tímarit | lid | . Slíkar málheildir eru sums staðar til |
| TIMARI | Títill: Talmál og málheildir – talmál og orðabækur | lid | sé ekki til hefur mikil árangur náðst |
| TIMARI | Höfundur: Ásta Svavarsdóttir | lid | býsks ritmáls með leitarmöguleikum) á Veraldarvefnum |
| TIMARI | Ritstjóri: Guðrún Kvaran | lid | (sbr. Sigrún Helgadóttir 2004) . |
| | Útgefandi: Stofnun Árna Magnússonar í íslenskum fræðum | | |
| | Í: Orð og tunga | | |
| | Ár: 2007 | | |

Oversigt over foredraget:

Hvor stammer projektet fra?

Hvad er et ordklasseopmærket korpus?

Hvordan bruges korpusset?

Oprettelse af MÍM korpusset

Tilgængelighed og brug

Andre korpusser

Hvor stammer projektet fra?

Det ordklasseopmærkede islandske korpus (MÍM) er et af de projekter som blev finansieret af Kultur- og undervisningsministeriets sprogteknologiske projekt 2000-2004.

Projektet blev startet i 2004 på *Orðabók Háskólans* og blev færdiggjort på *Stofnun Árna Magnússonar í íslenskum fræðum*

Anden finansiering:

- Nordisk Ministerråd (Nordisk Netordbog)
- Rannís (Viable Language Technology beyond English – Icelandic as a test case)
- De islandske studenters innovationsfond (nogle stipendier)
- Universitetets forskningsfond (nogle stipendier)
- META-NORD

Hvad er et ordklasseopmærket korpus?

Opmærket korpus er (e. *tagged corpus*)

- Samling af elektroniske tekster fra forskellige kilder som skal give indtryk om hvordan et sprog bliver brugt i en bestemt periode
- Hver enkelt tekst er forsynet med oplysninger om teksten (metadata), for eksempel titel, udgivelsesår, genre og forfatterens navn (eventuelt køn og fødselsår) for udgivne tekster
- Hvert ord er opmærket med oplysninger om ordklasse og bøjning
- Korpusset er gemt i standardiseret format (xml)

Hvordan bruges korpusset?

- I korpusset kan man finde information om:
 - frekvens af ordklasser, ord og bøjningsformer, om ordforbindelser, syntaks og semantik o.s.v.
- Er nyttigt når man laver for eksempel:
 - ordbøger, programmel til stave- og grammatikkontrol, maskinoversættelse, talegenkendelse og talesyntese, til støtte for handicappede (blinde, døve og hørehæmmede, bevægelseshæmmede og ordblinde) og i sprogundervisning

Oprettelse af MÍM

Formål:

- Samle tekster, skrevet af mennesker som har islandsk som modersmål, fra en række forskellige kilder med i alt 25 millioner ord fra perioden 2000–2009.
- Kun tekster som var elektronisk tilgængelige blev samlet
- Man skulle sikre licens til brug af teksterne i korpusset fra indehavere af ophavsret
- Alle tekster skulle opmærkes automatisk med oplysninger om ordklasse og bøjning

| Genre i MÍM | Antal tekst (filer) | Antal ord | % |
|---|---------------------------|-------------------|---------------|
| Trykte bøger | 168 | 5.972.893 | 23,89 |
| Aviser, trykte og elektroniske | 12.725 | 5.779.509 | 23,12 |
| Offentlige tekster (rapporter, domme, forslag, love, drøftelser fra Alþingi) | 1.246 | 3.513.990 | 14,06 |
| Tidsskrifter (trykte og elektroniske) | 311 | 2.501.222 | 10,00 |
| Blog | 8.998 | 1.976.706 | 7,91 |
| Artikler fra Universitetets videnskabsweb | 4.949 | 1.838.909 | 7,36 |
| Tekster fra websites for virksomheder, organisationer og institutter | 106 | 1.337.764 | 5,35 |
| Tekster som skal læses (blandt andet fra radio og tv) | 1.196 | 694.506 | 2,78 |
| Stile og skriftlige opgaver fra gymnasieelever og studenter | 51 | 666.042 | 2,66 |
| Talesprog | 4 | 504.318 | 2,02 |
| Usorteret | 46 | 214.663 | 0,86 |
| Total | 29.800 | 25.000.522 | 100,00 |

Oprettelse af MÍM

Tekstsamling

Tekst

- Tekst blev samlet fra trykte kilder og fra webben.
 - 58% trykte kilder
 - 40% fra webben
 - Ophavsretsbeskyttet tekst: 88,5%
- 2% talesprog, samlet i 4 forskellige projekter 2000–2006, 54 timer af transskriberet tale
 - Monologer (taler fra Alþingi)
 - Interviews
 - Spontane samtaler (2-5 personer)

Oprettelse af MÍM

Ophavsret

To juridiske dokumenter:

- Deklaration som indehavere af ophavsret underskriver
 - Teksten kan være tilgængelig uden betaling
 - Kun 80% af udgivet tekst inkluderes i korpusset
 - Teksterne er gjort tilgængelige med en brugerlicens
- Brugerlicens
 - Brugeren kan bruge sine resultater (det som han lærer fra korpusset) frit
 - Teksterne må ikke kopieres eller videregives til andre undtagen det som er antaget omfattet af citatretten

Oprettelse af MÍM

Forberedelse af tekst

Tekst blev fremskaffet i forskellige formater: pdf, xml, Word, tekst fra databaser, webtekst...

Uddrag af teksten

Fjerne fremmedsprogede og oldnordiske citater, fodnoter, indholdsfortegnelser, indekser, viser, tavler, billeder...

Fjerne bindestreg

Tekster blev fremskaffet enten i ISO-8859-1 eller UTF-8 tegnkodningstabel, alle tekster blev konverteret til UTF-8

Oprettelse af MÍM

Annotering - opmærkning

Segmentere i sætninger

Tokenisere - fordele i tokens

Bruge POS-tagger for at markere hvert ord i en tekst med oplysninger om ordklasse og morfologiske træk

Lemmatisere - beregne grundformen/opslagsformen af ord

Værktøjer: CorpusTagger som indeholder

IceNLP for segmentering og tokenisering

fire taggere for tagging (MXPOST, fnTBL, TriTagger, IceTagger)+CombiTagger som vælger tag

„Lemmald“ – for lemmatisering

Oprettelse af MÍM

Annotering - metadata

Tekstoplysninger (metadata)

Hver enkelt tekst er forsynet med oplysninger om tekstens oprindelse:

Udgivne tekster har bibliografiske oplysninger ligesom titel, udgivelsesår og genre og forfatterens navn (eventuelt køn og fødselsår)

Andre tekster har oplysninger som identificerer teksten

Metadata er vist via søgegrænsefladen og er en del af xml-filer som kan downloades.

Oprettelse af MÍM

Tilgængelighed

MÍM er tilgængelig på to forskellige måder:

1. Med søgning via søgesiden

<http://mim.arnastofnun.is/>

Brugere kan undersøge og se eksempler på sproglige fænomener sådan som de optræder i naturligt forekommende islandske tekster

2. Via „Download“ fra webstedet

<http://www.málföng.is/>

Er nyttig for dem som laver sprogteknologiske værktøjer

Oprettelse af MÍM

Tilgængelighed

1. Søgeseide

<http://mim.arnastofnun.is/>

Også tilgængelig gennem <http://arnastofnun.is/> og www.málföng.is.

Søgegrænsefladen bygger på Glossa (<http://www.hf.uio.no/tekstlab/glossa.html>) fra Universitetet i Oslo (Glossa bruger “corpus search engine”, IMS Corpus Workbench (CWB) fra Universitetet i Stuttgart (<http://cwb.sourceforge.net/>))

Oprettelse af MÍM

Tilgængelighed

2. „Download“

Teksterne er tilgængelige i TEI-konform xml-format i 29.800 filer fra webstedet

<http://www.málföng.is/>

Brugere må acceptere brugerlicens

Oprettelse af MÍM

Brug

1. Søgeseide

Brugere kan undersøge og se eksempler på sproglige fænomener sådan som de optræder i naturligt forekommende islandske tekster. Nyttig både for dem der beskæftiger sig professionelt med sprog (fx journalister, lærere og sprogforskere), og for dem der bare synes sprog er interessant og sjovt.

2. „Download“

Er nyttig for dem som laver sprogteknologiske værktøjer

Andre korpusser

To andre korpusser er tilgængelige på webstedet <http://mim.arnastofnun.is/>

1. Korpus for den islandske frekvensordbog

Indeholder omtrent 500.000 ord fra 100 tekster fra 1980–1889

Taggene blev korrigeret manuelt

Er velegnet til undervisning af morfologi

2. Saga korpusset

44 digitale tekster fra sagaer (41 islandske sagaer, Sturlunga, Heimskringla, Landnámabók) med omtrent 1.659.385 ord.

Nyttig til at undersøge brug af ord og konstruktioner i sagaerne

Søge i MÍM

MÍM kan være nyttigt til at se hvordan sproget bruges. Man kan tænke sig at det kan være nyttigt til dem som studerer islandsk som andet sprog, for elever og studenter på forskellige skoletrin og for dem der beskæftiger sig professionelt med sprog.

Vi skal undersøge tre eksempler på søgning i MÍM.

Søge i MÍM

1. Mange islandske sprogbrugere er ikke sikre på hvilken præposition („af“ eller „að“) bruges med „tilefni“ (anledning).

Man kan søge efter præposition som begynder med „a“ eventuelt efterfulgt af et ord efterfulgt af hvilken som helst bøjningsform af „tilefni“.

Søge i MÍM

2. Det kan være nyttigt at finde hvilket adjektiv bruges med forskellige sustantiver.

Vi kan for eksempel søge efter hvilke adjektiver bruges med ordet „kjóll“ (kjole).

Søg efter et adjektiv efterfulgt af hvilken som helst bøjningsform af „kjóll“.

Søge i MÍM

3. Det kan være ganske kompliceret i islandsk at finde ud hvilken præposition skal bruges med stednavne. Vi skal søge efter præpositionerne “í” (i) og „á“ (på) med stednavne sem har “fjörður” (fjord) som sidste led.

Søg efter „í“ eller „á“ efterfulgt af hvilken som helst bøjningsform af egennavn med “fjörður” som sidste led (stednavne i MÍM er taggede som egennavne)

MÍM

Tak for p  h  r!