Eiríkur Rögnvaldsson

Sprogteknologiske resourcer for islandsk leksikografi

Seminar om leksikografi og sprogteknologi

Schæffergården

31. januar 2010

Foredragets emne

- Islandsk sprogteknologi omkring århundredskiftet
- Sprogteknologiske resurser for islandsk
  - Morfologisk database
  - Balanceret tagget corpus
  - Sprogteknologiske værktøjer
- Open source policy
- Nye og kommende projekter

Islandsk sprogteknologi for ti år siden

- Islandsk sprogteknologi eksisterede ikke år 2000
- Vi havde
  - Et godt stavekontrolsystem
  - En primitiv talesyntese
- Vi havde ikke
  - Universitetsprogrammer eller kurser i sprogteknologi
  - Akademisk forskning i islandsk sprogteknologi
  - Software firmaer som arbejdede med sprogteknologi

Sprogteknologiudvalget

- Et specielt sprogteknologiudvalg
  - Oprettet i 1998 af ministeren for undervisning og forskning
- Opgaver:
  - At undersøge situationen i islandsk sprogteknologi
  - At komme med forslag for at styrke sprogteknologien i Island

Foreslåede aktiviteter

- Sproglige resurser skulle udvikles og opbygges
  - til andvendelse for fimaer som ville udvikle sprogteknologiske værktøjer og andre produkter
- Praktisk forskning indenfor sprogteknologi skulle støttes
- Firmaer skulle støttes til at udvikle sprogteknologiske produkter
- Universitetsprogrammer og kurser i sprogteknologi skulle oprettes

Sprogteknologiprogrammets produkter

- En morfologisk database med 258.000 ord
- Et balanceret tagget korpus med 25 million ord
- En statistisk PoS tagger
- Islandsk talesyntese
- Islandsk talgenkender
- Et forbedret stavekontrolsystem

Islandsk frekvensordbog

- 100 forskellige tekster fra 5 genrer
  - Islandske romaner
  - Romaner i oversættelse
  - Historie og biografier
  - Børnelitteratur
  - Faglitteratur
- 519.000 ord
  - 31.876 lemmaer

Fra Islandsk frekvensordbog

mælitæki *no*	**5**	**10**
mælitækið NHENG		1
mælitæki NHEÞ		1
mælitæki NHFN		3
mælitæki NHFO		1
mælitækin NHFOG		1
mælitækjum NHFÞ		3
mælskumaður *no*	**1**	**1**
mælskumaður NKEN		1
mælskur *lo*	**1**	**1**
mælskari LMVVEN		1
mæltur *lo*	**5**	**6**
mælt LFSHEO		2
mæltu LFSHEÞ		4

möglunarlaust *ao*	
möglunarlaust A-A	
mögulega *ao*	
mögulega A-A	
mögulegur *lo*	
mögulega LFSKFO	
möguleg LFSVEN	
mögulegar LFSVFO	
mögulegt LFSHEN	
mögulegt LFSHEO	
möguleg LFSHFN	
möguleg LFSHFO	
mögulegu LFVVEO	
mögulegu LFVVEE	

Fra basen til Islandsk frekvensordbog

• f p k e n	hann	han
• s f g 3 e þ	átti	ejede
• n h e o	afmæli	fødselsdag
• a o	í	i
• n k e o	dag	dag
• c	og	og
• n k e n g	hvolpurinn	hvalpen
• n k e n - m	Vaskur	Vask
• s f g 3 e þ	var	var
• n v e n	afmælisgjöf	fødselsdagsgave

Morfologisk database

- Morfologisk database for islandsk sprog
  - Beygingarlýsing íslensks nútímamáls, BÍN
- Et projekt der blev påbegyndt i 2002
  - ved Leksikografisk Institut
  - finanisieret af sprogteknologiprojektet
  - projektleder Kristín Bjarnadóttir
- Indeholder nu paradigmer for 258.000 ord
  - flere end 5,6 millioner ordformer

Hensigten med databasen

- Til hvilken brug blev databasen oprettet?
  - For brug indenfor sprogteknologi
  - For opslag på instituttets webside
- Har hidtil været brugt
  - i søgemaskiner (*embla* på mbl.is)
  - i telefonbogen
  - i læremateriale (Icelandic Online)
  - som hjælp ved tagging og lemmatisering

Hvad indeholder databasen?

Ordklasse	Lemmaer	Ordformer	Klasser
Substantiver	220.768	2.692.435	351
Verber	7.522	592.739	49
Adjektiver	25.779	2.339.466	31
Adverbier	1.979	2.239	29
Talord	78	1.845	2
Pronomener	42	820	
Artikel	1	24	

Paradigmer i BÍN-databasen

- Paradigmer for nogle ord
  - hestur subst.mask. ‘hest’
  - hvítur adj. ‘hvid’
  - bera vb. ‘bære’
  - inni adv. ‘inde’
  - þessi pron. ‘denne’
  - einn num. ‘én’

Omstrukturering af databasen

- Databasen er nylig blevet omstruktureret
  - filene lagt ind i en MySQL database
- Målet med omstruktureringen er
  - at gøre det nemmere at vedligeholde databasen
  - og at gøre søgning i den hurtigere
- Et excerperingsprogram er blevet lavet
  - i forbindelse med omstruktureringen

Balanceret korpus

- Balanceret tagget korpus
  - Projektleder Sigrún Helgadóttir
- 25 million ord
  - Mange forskellige teksttyper
- PoS-tagget
  - Samme tagsæt som i Islandsk frekvensordbog
- XML-markup
  - TEI-kompatibel format

Vigtigste teksttyper

- Avistekst
- Trykte bøger (romaner o.fl.)
- Blog
- Forskellige tidsskrifter
- Tekst fra Videnskabswebben
- Webtekster fra institutter, firmaer, etc.
- Love og andre tekster fra Altinget
- Talesprog

```
<s n=001>
<w type="fp1en" lemma="ég">ég</w> <w type="sfg1eþ" lemma="stökkva">stökk</w>
<w type="aa" lemma="á">á</w> <w type="aþ" lemma="eftir">eftir</w>
<w type="nkeþ" lemma="stræto">stræto</w> <w type="c" lemma="og">og</w>
<w type="sfg1eþ" lemma="veifa">veifaði</w> <p type="," lemma=",">,</p>
<w type="nkeng" lemma="vagnstjóri">vagnstjórinn</w> <w type="sfg3eþ" lemma="sjá">sá</w>
<w type="fp1eo" lemma="ég">mig</w> <w type="c" lemma="og">og</w>
<w type="sfg3eþ" lemma="stoppa">stoppaði</w> <p type="." lemma=".">.</p>
</s>
<s n=002>
<w type="fp1en" lemma="ég">ég</w> <w type="sfg1eþ" lemma="tauta">tautaði</w>
<w type="au" lemma="takk">takk</w> <w type="c" lemma="og">og</w>
<w type="sfg1eþ" lemma="brosa">brosti</w> <w type="ae" lemma="til">til</w>
<w type="fpkee" lemma="hans">hans</w> <w type="ao" lemma="um">um</w>
<w type="nveo" lemma="leið">leið</w> <w type="c" lemma="og">og</w>
<w type="fp1en" lemma="ég">ég</w> <w type="sfg1eþ" lemma="láta">lét</w>
<w type="nkeog" lemma="miði">miðann</w> <w type="sng" lemma="detta">detta</w>
<p type="." lemma=".">.</p>
</s>
<s n=003>
<w type="fpken" lemma="hann">hann</w>
<w type="sfg3eþ" lemma="láta">lét</w> <w type="nvfog" lemma="hönd">hendurnar</w>
<w type="sng" lemma="liggja">liggja</w> <w type="aa" lemma="fram">fram</w>
<w type="ao" lemma="á">á</w> <w type="nheog" lemma="stýri">stýrið</w>
<w type="c" lemma="og">og</w> <w type="sfg3eþ" lemma="horfa">horfði</w>
<w type="lkensf" lemma="þungbúinn">þungbúinn</w> <w type="aa" lemma="fram">fram</w>
<w type="ao" lemma="fyrir">fyrir</w> <w type="fpkeo" lemma="sig">sig</w>
<w type="aa" lemma="eins">eins</w> <w type="c" lemma="og">og</w>
<w type="fpken" lemma="hann">hann</w> <w type="svg3eþ" lemma="er">væri</w>
<w type="cn" lemma="að">að</w> <w type="sng" lemma="hugsa">hugsa</w>
<p type="." lemma=".">.</p>
</s>
```

Ordliste med lydskrift

- Liste med 56.000 frekvente ordformer
  - Oprindelig lavet for Hjal-projektet
- To typer af lydskrift
  - IPA og SAMPA
- Er allerede blevet brugt til
  - talgenkendelse
  - talesyntese

Fra den fonetiske ordliste

aðalatriði	a:ðalatrıðı	a:DaladrIDI
aðalatriðið	a:ðalatrıðıð	a:DaladrIDID
aðalatriðum	a:ðalatrıðʏm	a:DaladrIDYm
aðalástæðan	a:ðalaustaiðan	a:DalausdaiDan
Aðalbjörg	a:ðalpjœrk	a:Dalbj9rg
aðalbyggingu	a:ðalpıciŋkʏ	a:DalbIJ_iNgY
aðalbyggingunni	a:ðalpıciŋkʏnı	a:DalbIJ_iNgYnI
Aðaldal	a:ðaltal	a:Daldal
aðaldyrunum	a:ðaltı:rɔnʏm	a:DaldI:rOnYm
aðaleigandi	a:ðalei:ɣantı	a:Dalei:GandI
aðaleinkunn	a:ðaleiŋkʏn	a:DaleiN0gYn
aðaleinkunn	a:ðaleiŋkʏn	a:DaleiNkYn
aðalframkvæmdastjóri	a:ðalframkvaimtastjou:rı	a:Dalframkvaimdasdjou:rI
aðalfund	a:ðalfʏnt	a:DalfYnd
aðalfundar	a:ðalfʏntar	a:DalfYndar

Liste over verber med argumentstruktur

NF	draga	NH			
NF	draga	ÞF			
NF	draga	að	ÞF		
NF	draga	fram	ÞF		
NF	draga	frá	ÞF		
NF	draga	inn	ÞF		
NF	draga	niður	ÞF		
NF	draga	saman	ÞF		
NF	draga	ÞF	að		
NF	draga	ÞF	að	landi	
NF	draga	ÞF	að	sér	
NF	draga	ÞF	að	ÞGF	
NF	draga	ÞF	af	ÞGF	
NF	draga	ÞF	á	asnaeyrunum	
NF	draga	ÞF	á	eftir	sér
NF	draga	ÞF	áfram		

PoS taggere

Tagger	Unknown	Known	All
TnT	71.97	92.06	90.68
TnT*	73.10	92.85	91.50
IceTagger	75.36	92.95	91.76
Ice+HMM	75.70	93.20	92.01
BI+WC+CT	69.80	93.85	92.21
HMM+Ice	76.17	93.59	92.40
HMM+Ice+HMM	76.13	93.70	92.51

IceParser

Function type	F-measure using correct POS tags	F-measure using *IceTagger*	Freq. in test data
SUBJ	68.2%	47.6%	4.7%
SUBJ>	92.7%	89.4%	30.3%
SUBJ<	83.7%	75.1%	12.3%
OBJ	0.0%	0.0%	0.2%
OBJ>	43.5%	20.0%	0.8%
OBJ<	90.2%	78.2%	19.7%
OBJAP>	71.4%	57.2%	0.2%
OBJAP<	75.0%	46.2%	0.4%
OBJNOM<	30.8%	16.7%	0.6%
IOBJ<	73.3%	51.9%	0.9%
COMP	56.9%	40.0%	2.8%
COMP>	91.3%	91.3%	1.3%
COMP<	75.1%	70.0%	12.7%
QUAL	87.7%	77.9%	10.4%
TIMEX	74.7%	55.9%	2.7%
All	84.3%	75.3%	100.0%

Phrase type	F-measure using correct POS tags	F-measure using *IceTagger*	Freq. in test data
AdvP	91.8%	85.1%	8.2%
AP	95.1%	86.3%	8.1%
APs	87.0%	68.6%	0.5%
NP	96.8%	93.0%	37.6%
NPs	80.4%	74.3%	1.5%
PP	96.7%	91.3%	13.0%
VPx	99.2%	93.8%	19.3%
CP	100.0%	99.6%	5.7%
SCP	99.6%	97.6%	3.4%
InjP	100.0%	96.3%	0.2%
MWE	96.9%	92.6%	2.5%
All	96.7%	91.9%	100.0%

Lemmatisere

Lemmald	Tagged Input	Untagged Input
Basic (HOLI only)	97.85%	
+ Compound Analysis	98.38%	
+ Umlaut Substitution	98.42%	
+ Post processing	98.54%	
+ DMII	99.55%	
CST Lemmatizer		
	98.99%	93.15%

Online værktøjer

- IceNLP
  - IceTagger
  - IceParser
  - Lemmald
- http://nlp.cs.ru.is
  - værktøjerne kan bruges ét ad gangen
  - eller alle samtidig

Open source

- Sprogteknologiprojektets produkter skulle være tilgængelige for alle til en rimelig pris
  - men det har vist sig at selv en lav pris fører til at produkterne ikke bliver brugt
- Det er vigtigt at alle sprogteknologiske resurser for islandsk bliver open source
- IceNLP er blevet open source
  - http://sourceforge.net/projects/icenlp/
  - licenceret under GNU LGPL

Et nyt projekt

- Vi har nu startet et nyt treårigt projekt
  - for at opbygge resurser for islandsk sprogteknologi
- *Viable Language Technology beyond English*
  - *Icelandic as a test case*
- Tre delprojekter
  - En database af semantiske relationer
  - "Shallow transfer" maskinoversættelse
  - En træbank (syntaktisk analyseret korpus)

Konklusion

- Vi har fået en del sprogteknologiske resurser for islandsk i det sidste årti
- Nogen af dem vil også være vigtige for leksikografisk arbejde
- Vi mangler endnu en sprogteknologisk ord-database med morfologi, syntax og semantik
- Men vi har mange slags materiale som kunne kobles sammen til en sådan orddatabase

eirikur@hi.is