Automatiske metoder til excerpering af nye ord

Kristín Bjarnadóttir og Eiríkur Rögnvaldsson

Seminar om sprogrøgt og sprogteknologi
København 29. oktober 2007

HÁSKÓLI ÍSLANDS

Foredragets emne

- Prototype af et værktøj for automatisk excerpering af nye ord i islandsk
  - Work in progress
- Oversigt
  - En kort beskrivelse af morfologidatabasen
  - Excerperingsprocessen
  - Konklusion

Den morfologiske database

- Morfologisk database for islandsk sprog
  - Beygingarlýsing íslensks nútímamáls, BÍN
- Et projekt der blev påbegyndt i 2002
  - finanisieret af islandsk sprogteknologifond
  - projektleder Kristín Bjarnadóttir
- Indeholder nu paradigmer for 258.000 ord

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 lige blevet omstruktureret
  - af Hjálmar Gíslason og Kristín Bjarnadóttir
- Målet med omstruktureringen er
  - at gøre det nemmere at vedholde databasen
  - og at gøre søgning i den hurtigere
- Et excerperingsprogram er blevet lavet
  - i forbindelse med omstruktureringen

Excerperingsprocessen

- Excerperingsprogrammet
  - læser teksten (enten direkte fra en fil eller tekst som har været kopieret eller tastet ind på excerperingssiden)
  - slår hvert enkelt ord op i databasen
  - skriver ud en liste over alle ordformer som ikke findes i databasen

Demonstration af excerperingen

- For at demonstrere excerperingen vil vi
  - gå til den største islandske netavis, mbl.is
  - vælge en nyhed fra forsiden
  - gå til excerperingssiden og kopiere teksten ind
  - klikke på “orðtaka” (excerpering)
  - få frem en liste af nye (ukendte) ord i teksten

Prøvetekst 1

- Fra den islandske original af foredraget:
  - Samkeyrsla við uppflettiorð í BÍN; bæta þarf inn efni úr Orðabankanum til þess að vit verði í útkomunni úr orðtökutólinu m.t.t. nýyrða. Finna þarf upp fljótvirka leið til að keyra saman flettulistana. E.t.v. er hægt að fara bara aðra umferð í orðtökutólinu?

Rengøring og analyse

- Fjernelse af fejl og udenlandske ord
  - kan gøres halvautomatisk
- Lemmatisering
  - endnu manuel men bliver automatisk
- Morfologisk analyse
  - udarbejdet i forbindelse med lemmatisering

Resultat af processen

- Endeligt resultat af hele processen
  - *Ordliste* *Lematisering+analyse*
  - Orðabankanum Orða.banki
  - orðtökutól orðtöku.tól
  - flettulistana flettu.listi
- Hvad slags ord er disse tre?
  - er de virkelige nye ord eller huller i BÍN?

Prøvetekst 2

- Blogtekst fra internettet
  - 1.774 ord
  - skrevet af en ung kvindelig student
  - indeholder en del udenlandske ord
  - både tilpassede og utilpassede
  - kendte ord skrives sædvanligvis korrekt
- Kig på hele teksten

Resultat af excerperingen

- Excerperingen giver ialt 75 nye lemmaer
  - Nye ord som kan betragtes som islandsk 30
  - Retskrivningsvarianter 3
  - Forældede ord 1
  - Uforståelige former 5
  - Udenlandske ord 36
    - (engelsk 24, fransk 2, latin 5, spansk 1, ? 4)

Ordtyper efter rengøring og analyse

- Huller i BÍN – ord som ikke er nye i sproget
  - og skal absolut føjes til BÍN
- Produktive sammensætninger
  - som ikke nødvendigvis skal tilføjes
- Nye ord
  - neologismer, låneord

Typer af nye ord - 1

- *Huller i BÍN*
- baksveifla
- hraðaspurning
- menntaskólaár
- oggu
- sandkastali
- skólafélag
- trélitur

- *Produktive sammensætninger - skal de komme med?*
- laufblaðahrúga
- smáslagsmál
- útúrgeldur
- þágufallsnotkun
- þágufallsþjáningarsystir

Typer af nye ord - 2

- *Nye ord?*
- hollívúdkoss
- Kakóland
- samnörd
- tískumógúll
- týpuföt
- týpuskapur
- *"Slettur" (pletter)*
- attitjút
- cirkabát
- fokking
- gúgl
- gúgla
- sushi

Ord i kontekst

- Efter valg af ord kører vi et program som
  - læser den oprindelige tekst igen
  - finder de ord som vi har valgt
  - skriver ordene i kontekst ud i en fil
  - med information om tekst, side osv.

Hvilken nytte har vi af de nye ord?

- De nye ord kan bruges som
  - Tilføjelser til BÍN
  - Materiale til ordbogsarbejde
  - Materiale til terminologilister
  - Materiale for lingvister og sprogteknologer
  - Materiale for sprogrøgtere

Værktøjet i sprogrøgten

- Hvordan kan man bruge dette værktøj
  - for at følge med udvikling og ændringer i ordforrådet?
- Forslag: Kontakte alle islandske aviser
  - få dem til at køre excerperingsværktøjet daglig pa alle deres tekster
- Dette ville give os en mængde nye ord

Konklusion

- Denne metode giver os en masse nye ord
  - egentlige neologismer
  - huller i BÍN
  - produktive sammensætningar
  - varianter i retskrivning og orddannelse
- Det tager væsentlig tid at rense listen
  - men det kunne gøres halvautomatisk